Fundamental Limits of Cloud and Cache-Aided Interference Management with Multi-Antenna Edge Nodes

Jingjing Zhang and Osvaldo Simeone

Abstract

In fog-aided cellular systems, content delivery latency can be minimized by jointly optimizing edge caching and transmission strategies. In order to account for the cache capacity limitations at the Edge Nodes (ENs), transmission generally involves both fronthaul transfer from a cloud processor with access to the content library to the ENs, as well as wireless delivery from the ENs to the users. In this paper, the resulting problem is studied from an information-theoretic viewpoint by making the following practically relevant assumptions: 1) the ENs have multiple antennas; 2) only uncoded fractional caching is allowed; 3) the fronthaul links are used to send fractions of contents; and 4) the ENs are constrained to use one-shot linear precoding on the wireless channel. Assuming offline proactive caching and focusing on a high signal-to-noise ratio (SNR) latency metric, the optimal information-theoretic performance is investigated under both serial and pipelined fronthaul-edge transmission modes. The analysis characterizes the minimum high-SNR latency in terms of Normalized Delivery Time (NDT) for worst-case users' demands. The characterization is exact for a subset of system parameters, and is generally optimal within a multiplicative factor of $3/2$ for the serial case and of 2 for the pipelined case. The results bring insights into the optimal interplay between edge and cloud processing in fog-aided wireless networks as a function of system resources, including the number of antennas at the ENs, the ENs' cache capacity and the fronthaul capacity.

Index Terms

Fog, cloud, cellular system, edge caching, interference management.

I. Introduction

Content delivery is one of the most important use cases for mobile broadband services in 5G networks. A key technology that promises to help minimize delivery latency and network congestion is edge caching, which relies on the storage of popular contents at the ENs, i.e., at the base stations or access points. Initial works on the subject [1] studied the advantages of edge caching in terms of cache hit probability, hence adopting the standard performance criteria used in the networking literature (see e.g., [2]). The information-theoretic analysis of edge caching, which has been undertaken in the past few years starting with [3], has instead concentrated on the impact of the cached content distribution across the ENs on the ENs' capability to carry out interference management (see also [4]). As a general observation, caching the same content across multiple ENs

The authors are with the Department of Informatics, King's College London, London, UK (emails: jingjing.1.zhang@kcl.ac.uk, osvaldo.simeone@kcl.ac.uk).

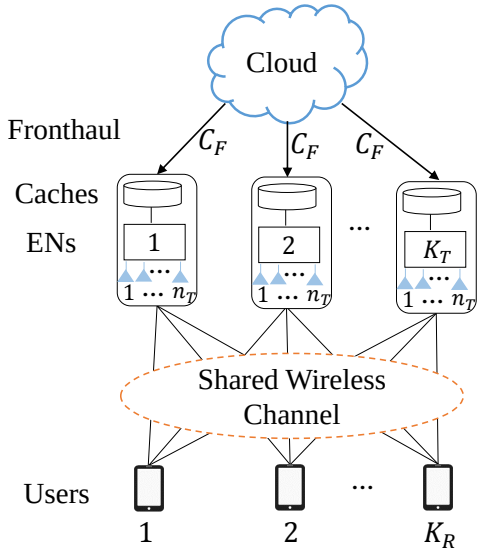

Fig. 1. Cloud and cache-aided F-RAN system model with multi-antenna ENs.

enables cooperative delivery strategies involving multiple ENs, whereas properly placed distinct contents can yield coordination opportunities [3]. The relative effect of interference management via coordination or cooperation on content delivery is best studied in the high-SNR regime, in which the performance is limited by interference, as done in [3], [5].

In most existing works, as further discussed below, the high-SNR analysis of the interference management capabilities of cache-aided systems was performed under the assumption that the overall cache capacity available in the system, including at the users, is sufficient to store the entire library of popular contents. When this assumption is violated, contents need to be retrieved from a content server by leveraging transport links that connect the ENs to the access or core networks. This more general scenario was first studied from an information-theoretic perspective in [6], [7]. In these works, a cloud processor is assumed to be connected to the ENs via so called fronthaul links, as seen in Fig. 1.

For the model in Fig. 1, which is referred to as Fog-Radio Access Network (F-RAN), the key design problem concerns the optimal use of fronthaul and wireless edge resources for caching and delivery. Assuming the standard offline caching scenario with static popular set, reference [7] identified high-SNR optimal caching and delivery strategies within a multiplicative factor of two. The approximately optimal scheme in [7] relies on both Zero-Forcing (ZF) one-shot precoding and interference alignment for transmission on the wireless edge channel and on cloud precoding and on quantization [8]. In this work, we revisit the results in [7] by making the following practically relevant assumptions: 1) the ENs have multiple antennas; 2) only uncoded fractional caching is allowed; 3) the fronthaul links can only be used to send uncoded fractions of contents; and 4) the ENs are constrained to use linear precoding on the wireless channel.

Related Work: Assuming offline caching, cache-aided interference management was first studied in [3], in which transmitter-side caches are considered, and a delivery strategy is proposed by leveraging interference alignment, ZF precoding and interference cancellation. Extensions that account for caching at both transmitter and receiver sides were provided in [5], [9]–[12]. In [10], a novel strategy based on the separation of physical and network layers is investigated. Under the assumption of one-shot linear precoding, references [5], [12] reveal that the transmitters' caches and receivers' caches contribute equally to

the high-SNR performance. More general lower bounds were derived in [13] (see also [7]), for which matching upper bounds were found in [6], [14] in special cases. Decentralized caching is studied in [15]. In [16], [17], the performance of edge caching with partial connectivity is studied without any channel state information (CSI), while imperfect CSI is considered in [18], [19].

The joint design of cloud processing and edge caching for the F-RAN model was studied in references [6], [7] and then in [20]–[22], by focusing on the high-SNR latency performance metric known as Normalized Delivery Time (NDT) proposed in [7]. This metric essentially measures the inverse of the number of degrees of freedom (DoF) [6], [7]. Reference [7] characterizes the minimum NDT with a multiplicative factor of 2 by considering single-antenna ENs, intra-file coding for caching and general delivery strategies on fronthaul and edge channels. References [20], [23] studied a related scenario with coexisting macro- and small-cell base stations. The work [24] studied an F-RAN model with decentralized placement algorithms. In contrast to the abovementioned works, references [22], [25] investigated online caching in the presence of a time-varying set of popular files. Overall, the set-up studied in this work extends the model in [5] by including cloud processing, fronthauling, and multi-antenna ENs, but, unlike [5], it excludes caching at the receivers' sides.

Main contributions: This paper investigates interference management in a cloud and cache-aided F-RAN model, as illustrated in Fig. 1, with multiple antennas at the transmitters, under the assumptions of one-shot linear precoding and transmission of uncoded contents on the fronthaul links. The NDT is adopted as a high-SNR delivery latency metric. The main contributions are summarized as follows.

- We first derive upper bounds on the minimum NDT under the assumption of serial fronthaul-edge transmission when only edge caching or only fronthaul resources (and no cache capacity) are available for delivery. In the serial delivery mode, fronthaul transmission is followed by edge transmission. The proposed schemes use clustered ENs' cooperation via ZF beamforming to cancel interference on the wireless edge channel. Cooperation is enabled by contents shared thanks to edge caching during placement phase, or by fronthaul transmissions in the delivery phase. The caching and fronthauling strategies rely on an efficient packetization method that is separately at most linear in the number of transmitters and receivers.
- For the general F-RAN set-up, an upper bound on the minimum NDT is derived as a function of the cache storage capacity, the fronthaul rate, and the number of ENs' antennas. To this end, we propose a caching and delivery scheme that manages interference via ZF by means of the ENs' clustered cooperation as enabled by both fronthaul and edge caching resources. Then, an information-theoretic lower bound on the minimum NDT is derived. As a result, the minimum NDT is characterized exactly for a large subset of system parameters, and approximately within a multiplicative factor of $3/2$ for any value of the parameters.
- We finally study a pipelined delivery mode whereby fronthaul and edge transmissions can take place simultaneously. We show that the NDT under pipelined transmission can be improved as compared to serial delivery, and the minimum NDT is derived within a multiplicative factor of 2.

The rest of the paper is organized as follows. Section II describes a general $K_R \times K_T$ F-RAN model and the NDT performance metric for serial fronthaul-edge delivery. Section III and Section IV study the specific set-ups of edge-only and fronthaul-only

F-RAN, respectively, while the general F-RAN set-up is investigated in Section V. Under serial transmission, the upper bounds and the proposed scheme are presented along with a finite optimality gap from the minimum NDT. The pipelined fronthaul-edge transmission is discussed in Section VI, where upper and lower bounds on the minimum NDT are presented. Section VII concludes the work and also highlights future research directions.

Notation: For any integer K, we define the set $[K] \triangleq \{1, 2, \cdots, K\}$. For a set A, $|A|$ represents the cardinality. We use the notation $\{f_n\}_{n=1}^{N} \triangleq \{f_1, \cdots, f_n, \cdots, f_N\}$. For function $g(n)$, the notation $f(n) = o(g(n))$ denotes a function $f(n)$ that satisfies the limit $\lim_{n\to\infty}(f(n)/g(n)) = 0$. The ceiling function $\lceil x \rceil$ maps x to the least integer that is greater than or equal to x, and the floor function $\lfloor x \rfloor$ maps x to the greatest integer that is less than or equal to x. Moreover, the nearest positive integer function $[x]$ returns the nearest positive integer to x. We also have $(x)^+ \triangleq \max\{x, 0\}$.

II. System Model and Performance Metric

In this section, we present the model under study, which consists of an F-RAN system with multi-antenna ENs that performs the hard transfer of uncoded contents on the fronthaul links and one-shot linear precoding on the wireless edge channel. We also adapt the NDT metric [7] to this model. We consider the serial mode of delivering across fronthaul and edge channels, and discuss the pipelined mode in Section VI.

A. System Model

We consider the F-RAN model shown in Fig. 1 where a set $\mathcal{K}_T = \{1, \cdots, K_T\}$ of ENs, each having n_T antennas, are connected to K_R single-antenna receivers $\mathcal{K}_R = \{1, \cdots, K_R\}$ through a shared wireless channel, as well as to a cloud processor (CP) via fronthaul links. The CP has access to a library of N files $\{W_n\}_{n=1}^{N}$, of L bits each. Any file W_n contains F packets $\mathcal{W}_n = \{W_{nf}\}_{f=1}^{F}$, where each packet W_{nf} is of size L/F bits, and F is an arbitrary parameter. Note that we refer to the set of packets $\{W_{nf}\}_{f=1}^{F}$ in file W_n as $\mathcal{W}_n$. Each fronthaul link has capacity C_F bits per symbol, where a symbol refers to a channel use of the wireless channel, and each EN has a cache with capacity of μNL bits, with $\mu \in [0, 1]$. Parameter μ is referred to as the fractional cache size.

In the *pre-fetching phase*, the caches of the ENs are pre-filled with content from the library under the cache capacity constraints. The content of the cache of each EN i is described by the set $\mathcal{C}_i = \{\mathcal{C}_{i1}, \cdots, \mathcal{C}_{in}, \cdots, \mathcal{C}_{iN}\}$, where $\mathcal{C}_{in} \subseteq \mathcal{W}_n$ represents the subset of packets from file W_n that are cached at EN i. Due to the cache capacity constraint, its size must satisfy the inequality

$$\frac{|\mathcal{C}_{in}|}{F} \leq \mu. \tag{1}$$

Note that as in [5], the model at hand allows for no coding of the cached content either within or across files.

In the *delivery phase*, each user k requests a file W_{d_k}, with $d_k \in [N]$, from the library. Given the request vector $\boldsymbol{d} = \{d_1, \cdots, d_{K_R}\}$ and the CSI on the edge channel, to be discuss below, the CP transmits information about the requested files $\{W_{d_1}, \cdots, W_{d_{K_R}}\}$ to the ENs via the fronthaul links. Specifically, on each fronthaul i, the set $\mathcal{F}_i = \{\mathcal{F}_{id_1}, \cdots, \mathcal{F}_{id_{K_R}}\}$ of packets is sent, where $\mathcal{F}_{id_k} \subseteq \mathcal{W}_{d_k}$ is a subset of packets from file W_{d_k}. Note that, as mentioned, the described model assumes hard-transfer fronthauling of uncoded packets. After the fronthaul transmission, any EN i has access to the fronthaul information

$\mathcal{F}_i$, as well as to the cached content $\mathcal{C}_i$. This information is used by the ENs to deliver the users' requests $\{W_{d_1}, \cdots, W_{d_{K_R}}\}$ through the wireless channel.

To this end, we constrain the wireless transmission strategy to one-shot linear precoding by following [5]. Accordingly, wireless transmission takes place over B blocks to deliver the $K_R F$ desired packets. In any block $b \in [B]$, the ENs send a subset of the requested packets, denoted by $\mathcal{D}(b) \subseteq \{W_{d_1 f}, \cdots, W_{d_K f}\}_{f=1}^{F}$, to a subset $\mathcal{R}(b)$ of K_T users, such that each user in $\mathcal{R}(b)$ can decode exactly one packet without interference by the end of the block. To this purpose, in any block b, each EN i sends a linear combination of the subset of the packets in $\mathcal{D}(b)$ that it has access to in its cache, i.e., in $\mathcal{C}_i$, or that it has received on the fronthaul link, i.e., in $\mathcal{F}_i$. For any given symbol within the block, the transmitted signal of EN i is hence given as

$$\boldsymbol{x}_i(b) = \sum_{\substack{(n,f):\\ W_{nf} \in \mathcal{D}(b) \cap \{\mathcal{C}_i \cup \mathcal{F}_i\}}} \boldsymbol{v}_{inf}(b) s_{nf}(b), \tag{2}$$

where $s_{nf}(b)$ is a coded symbol for packet W_{nf}, and $\boldsymbol{v}_{inf}(b) \in \mathbb{C}^{n_T \times 1}$ is the precoding vector for the same file. As we have described, each packet $W_{nf} \in \mathcal{D}(b) \cap \{\mathcal{C}_i \cup \mathcal{F}_i\}$ is intended for a single user in $\mathcal{R}(b)$. We impose the power constraint $\mathbb{E}[||\boldsymbol{x}_i(b)||^2] \leq P$.

The received signals of each user $k \in \mathcal{R}(b)$ in block b is given as

$$y_k(b) = \sum_{i=1}^{K_T} \boldsymbol{h}_{ki}^T(b) \boldsymbol{x}_i(b) + z_k(b), \tag{3}$$

where $\boldsymbol{h}_{ki}(b) \in \mathbb{C}^{n_T \times 1}$ is the channel vector between EN i and user k, and $z_k(b)$ is the zero-mean complex Gaussian noise with normalized unitary power. The channels $\{\boldsymbol{h}_{ki}(b)\}_{k \in [K_R], i \in [K_T], b \in [B]}$ are arbitrary as long as the set $\{\boldsymbol{h}_{ki}(b)\}_{k \in [K_R], i \in [K_T]}$ is linearly independent for each block b. In each block b, we assume that all the ENs and users have access to the full CSI $\{\boldsymbol{h}_{ki}(b)\}_{k \in [K_T], i \in [K_R]}$ as necessary. The delivery of the packets in the set $\mathcal{D}(b)$ is said to be achievable if there exist precoding vectors $\{\boldsymbol{v}_{inf}(b)\}$, such that, with full CSI, each user $k \in \mathcal{R}(b)$ can decode without interference its intended packet. Given that the users have a single antenna, this happens if the received signal $y_k(b)$ is directly proportional to the desired symbol $s_{nf}(b)$ plus additive Gaussian noise with constant power, i.e., not scaling with the signal power P. The resulting point-to-point interference-free channel from the ENs to each served user k supports transmission at rate $\log(P) + o(\log(P))$.

B. Performance Metric: NDT

Consider a given policy defined by the parameters $\{\mathcal{C}_i, \mathcal{F}_i, \{\boldsymbol{v}_{inf}(b)\}_{n \in [N], f \in [F], b \in [B]}\}_{i=1}^{K_T}$, where the fronthaul messages $\{\mathcal{F}_i\}_{i=1}^{K_T}$ and the beamforming vectors $\{\boldsymbol{v}_{inf}(b)\}_{n \in [N], f \in [F], b \in [B], i \in [K_T]}$ are defined on request vector $\boldsymbol{d}$ and CSI $\{\boldsymbol{h}_{ki}(b)\}_{k \in [K_R], i \in [K_T], b \in [B]}$. Given the fronthaul messages defined by the subsets $\{\mathcal{F}_i\}_{i=1}^{K_T}$, the time required for fronthaul transmission can be computed as

$$T_F = \max_{i \in [K_T]} \frac{|\mathcal{F}_i| L}{F} \frac{1}{C_F}, \tag{4}$$

since $|\mathcal{F}_i|$ packets with $|\mathcal{F}_i| L/F$ bits need to be delivered to EN i over a fronthaul link of capacity C_F and T_F is the maximum among the K_T fronthaul latencies. Furthermore, given the delivered packet set $\{\mathcal{D}(b)\}_{b=1}^{B}$, the total time needed for wireless

edge transmission over B blocks is

$$T_E = \frac{BL}{F} \frac{1}{(\log(P) + o(\log(P))}. \tag{5}$$

This is because, in each of the B blocks, one packet with L/F bits is sent to each user in $\mathcal{R}(b)$ at rate $\log(P) + o(\log(P))$.

As in [7], we normalize the latency by the term $L/\log(P)$. This corresponds to the transmission latency, neglecting $o(\log(P))$ terms, for a reference system that transmits interference-free to all users at the maximum rate $\log(P)$. Moreover, as in [7], we evaluate the impact of the fronthaul capacity C_F in the high-SNR regime by using the scaling $C_F = r\log(P)$, so that the parameter r measures the ratio between the fronthaul capacity and the interference-free wireless channel capacity to any user. Accordingly, we define the fronthaul NDT of the given policy as

$$\delta_F = \lim_{P\to\infty} \lim_{L\to\infty} \frac{T_F}{L/\log(P)} = \max_{i\in[K_T]} \frac{|\mathcal{F}_i|}{Fr}, \tag{6}$$

and the edge NDT as

$$\delta_E = \lim_{P\to\infty} \lim_{L\to\infty} \frac{T_E}{L/\log(P)} = \frac{B}{F}. \tag{7}$$

Assuming serial fronthaul and edge transmission, the overall NDT is given as

$$\delta = \delta_E + \delta_E. \tag{8}$$

For any pair (μ, r), the minimal NDT across all achievable policies $\{\mathcal{C}_i, \mathcal{F}_i, \{\boldsymbol{v}_{inf}(b)\}_{n\in[N],f\in[F],b\in[B]}\}_{i=1}^{K_T}$ is defined as

$$\bar{\delta}(\mu, r) = \inf\{\delta(\mu, r) : \delta(\mu, r) \text{ is achievable for some } F \geq 1\}. \tag{9}$$

Note that in the definition (9), we allow for a partition of the files in an arbitrary number of F packets. By construction, we have the inequality $\bar{\delta}(\mu, r) \geq 1$, where the lower bound is achieved in the mentioned ideal system. By allowing for time sharing among different policies, we finally define the minimum NDT as

$$\delta^*(\mu, r) = \text{l.c.e.}(\bar{\delta}(\mu, r)). \tag{10}$$

where the lower convex envelope (l.c.e.)[1] is computed throughout this paper by considering $\bar{\delta}(\mu, r)$ as a function of μ. The achievability of $\delta^*(\mu, r)$ given the achievable NDT $\bar{\delta}(\mu, r)$ follows by a standard cache and time-sharing argument, which is detailed in [7, Lemma 1].

III. Achievable NDT for Edge-only Caching

As a preliminary result to be leveraged in Section V, here we describe an achievable NDT (Proposition 1) and the corresponding caching and delivery scheme for the described F-RAN model with edge-caching only, i.e., with zero fronthaul rate $(r = 0)$. Note that in this regime, the condition $\mu \geq 1/K_T$ needs to be satisfied in order to ensure a finite NDT. In fact,

[1]The l.c.e, is the supremum of all convex functions that lie under the given function.

otherwise, contents could not be fully cached across the ENs. The proposed scheme uses clustered ZF cooperation as in [5] but via a more efficient packetization method.

A. Achievable NDT

In the proposed scheme, in each block b, a cluster of ENs serve a given number u of users by using cooperative ZF precoding on the wireless channel. Cooperation at the ENs via ZF is enabled by the availability of shared contents across the caches of the ENs. To quantify the content availability at the ENs, we define the multiplicity m of any file as the number of times that the file appears across all the ENs. Via edge caching, a multiplicity $m(\mu) = \lfloor \mu K_T \rfloor$ can be ensured for all contents by edge caching since μK_T is the per-file cache capacity across all the ENs. Given the multiplicity m, it will be shown that contents can be allocated so that clusters of m ENs can transmit cooperatively in each block to serve up to mn_T users on interference-free channels via ZF beamforming. Note that mn_T is in fact the total number of transmit antennas available at a cluster of m ENs. Hence, the number of users that can be served via ZF in each block is given as

$$u(m) = \min\{mn_T, K_R\}. \tag{11}$$

Note that, by (11), the multiplicity m of a content can be upper bounded without loss of optimality by

$$m_{max} = \min\left\{K_T, \left\lceil \frac{K_R}{n_T} \right\rceil \right\}. \tag{12}$$

This is because, when the multiplicity reaches m_{max}, the ENs can cooperate in each block via ZF beamforming to completely eliminate inter-user interference for the maximum number of users, which is given by $\min\{K_T n_T, K_R\}$. The resulting achievable NDT is presented in the following proposition.

Proposition 1: For an F-RAN system with any cache capacity $\mu \in [1/K_T, 1]$ and fronthaul rate $r = 0$, we have the upper bound on the minimum NDT $\delta^*(\mu, r = 0) \leq \delta_E(m(\mu))$, where we have defined the edge NDT as a function of the multiplicity m as

$$\delta_E(m) = \frac{K_R}{u(m)}, \tag{13}$$

with function $u(m)$ in (11), and the multiplicity

$$m(\mu) = \min\{\lfloor \mu K_T \rfloor, m_{max}\}. \tag{14}$$

Proof: The proof is reported in Section III-C. ■

B. Examples

Before proving a sketch of proof, we discuss two examples that illustrate the achievable cache-aided delivery scheme. We consider an F-RAN model with $r = 0, K_T = 4$ ENs and $n_T = 2$ per-EN antennas (see Fig. 2).

Example 1. Consider $K_R = 4$ and $\mu = 0.5$, so that, by (14), we have the multiplicity $m(\mu) = 2$ in (14). Each library file W_n is divided into $F = 2$ equal and disjoint packets $\{W_{n1}, W_{n2}\}$. As illustrated in Fig. 2(a), in the caching phase, the

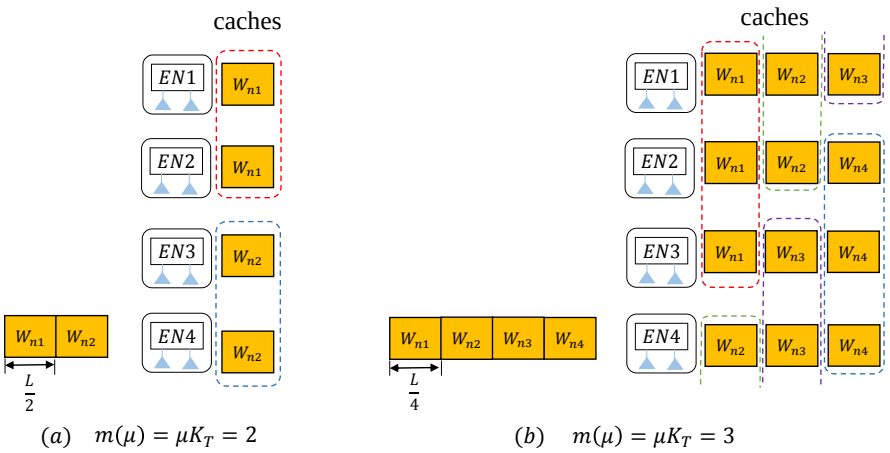

Fig. 2. Examples of caching scheme under edge-only transmission for the achievable NDT in Proposition 1. The dashed lines identify the clusters of cooperative ENs in (18).

ENs are divided into two clusters of $m(\mu) = 2$ ENs each: the first cluster, consisting of EN 1 and 2, caches the first packets $\{W_{n1}\}_{n=1}^{N}$, while the second cluster, consisting of EN 3 and 4, caches the second packets $\{W_{n2}\}_{n=1}^{N}$. As a result, we have the subsets $\mathcal{C}_1 = \mathcal{C}_2 = \{W_{n1}\}_{n=1}^{N}$, and $\mathcal{C}_3 = \mathcal{C}_4 = \{W_{n2}\}_{n=1}^{N}$. In the delivery phase, for any demand vector $\boldsymbol{d}$, the ENs in the first cluster can send packets $\{W_{d_k1}\}_{k=1}^{4}$ cooperatively via ZF to all the $u(m) = K_R = 4$ users at a time in one block, since the cluster collectively has four antennas. In a similar manner, packets $\{W_{d_k2}\}_{k=1}^{4}$ can be delivered by EN 3 and 4 to all the users in one block. The resulting NDT in (7) is $\delta_E = B/F = 2/2 = 1$.

Example 2. Consider now a cache capacity $\mu = 3/4$. In this case, the multiplicity (14) equals $m(\mu) = 3$, which is not a divisor of K_T. This requires a more complex placement strategy that accounts for the need to define clusters of $m(\mu) = 3$ cooperative ENs. To this end, in the proposed scheme, for caching, each file W_n is split into $F_C = 4$ disjoint parts of equal size, i.e., $W_n = \{W_{ni}\}_{i=1}^{4}$. As seen in Fig. 2(b), the caching policy places each part W_{ni} at a contiguous cluster of $m(\mu) = 3$ ENs, where contiguity is defined in a circular manner with respect to the EN index in set $\mathcal{K}_T$. More concretely, the ENs are clustered into four subsets, defined as $\mathcal{K}_{T1} = \{1, 2, 3\}$, $\mathcal{K}_{T2} = \{4, 1, 2\}$, $\mathcal{K}_{T3} = \{3, 4, 1\}$, and $\mathcal{K}_{T4} = \{2, 3, 4\}$. For any popular file W_n, part W_{ni} is placed at all ENs in $\mathcal{K}_{Ti}$ for $i = 1, 2, 3, 4$.

Consider the worst-case request of K_R distinct files. The clusters $\{\mathcal{K}_{Ti}\}_{i=1}^{4}$ of ENs are activated in turn to transmit all the requested parts $\{W_{d_ki}\}_{i=1}^{4}$ to all $k \in \mathcal{K}_R$ user. Since each EN has $n_T = 2$ antennas, with EN cooperation among the three ENs in each cluster, up to $mn_T = 6$ users can be served simultaneously in each block. If K_R is smaller than $mn_T = 6$ or a multiple thereof, it is hence possible to serve groups of $u(m) = \min\{mn_T, K_R\}$ distinct users in each block.

Suppose now instead that we have $K_R = 8$, which does not satisfy this condition. In a manner similar to the definition of the clusters of ENs, we define $B_D = 4$ groups of six users $\mathcal{K}_{R1} = \{1, 2, 3, 4, 5, 6\}$, $\mathcal{K}_{R2} = \{7, 8, 1, 2, 3, 4\}$, $\mathcal{K}_{R3} = \{5, 6, 7, 8, 1, 2\}$, and $\mathcal{K}_{R4} = \{3, 4, 5, 6, 7, 8\}$. In order to serve $u(m) = 6$ users simultaneously in each block, each part W_{d_ki} of a requested file W_{d_k} is further split into $F_D = 3$ equal packets as $W_{d_ki} = \{W_{d_kij}\}_{j=1}^{3}$. For any EN cluster $\mathcal{K}_{Ti}, i \in [F_C]$, the ENs can cooperatively send a subset of $u(m) = 6$ packets from $\{W_{d_ki}\}_{k=1}^{K_R}$ to all users in group $\mathcal{K}_{Rj}$ when $j = 1, \cdots, B_D$, requiring B_D blocks. As a result, we have $B = F_C B_D = 16$ blocks and $F = F_C F_D = 12$ packets, yielding the NDT $\delta_E = B/F = 4/3$.

C. Proof of Proposition 1

We now generalize the proposed scheme. For any cache capacity μ, by (14), we have the multiplicity $m = m(\mu) = \min\{\lfloor \mu K_T \rfloor, m_{max}\}$. To start, we define the number of parts used during the caching phase as

$$F_C = \frac{\text{l.c.m.}(m, K_T)}{m}, \tag{15}$$

where l.c.m.(a, b) is the least common multiple of integers a and b. As we will see, this choice guarantees that clusters of m ENs can store the same part of each file, enabling cooperative transmission. We also define the number of packets created out of each cached part as

$$F_D = \frac{\text{l.c.m.}(u(m), K_R)}{K_R}. \tag{16}$$

This ensures that in each block, subsets of $u(m)$ users can be served. Overall, the number of packets is

$$F = F_C F_D. \tag{17}$$

By (17), we have the inequality $K_T/m \leq F \leq K_T K_R$, where the lower bound is attained when K_T and K_R are divisible by m and $u(m)$, respectively. The upper bound demonstrates that the proposed scheme requires a packetization into a number of packets no larger than the product $K_T K_R$. This stands in contrast to the method of [5], which, when specialized to the case of no caching at the receivers, requires a number of packets that is exponential in the number of transmitters.

In the caching phase, each file W_n is equally split into F_C parts $\{W_{ni}\}_{i=1}^{F_C}$. Correspondingly, the ENs are clustered into F_C clusters, defined as $\{\mathcal{K}_{Ti}\}_{i=1}^{F_C}$[2], where each cluster is defined as

$$\mathcal{K}_{Ti} = \{[(i-1)m + 1]_{K_T}, [(i-1)m + 2]_{K_T}, \cdots, [im]_{K_T}\}, \tag{18}$$

where $[a]_b = 1 + \text{mod}(a - 1, b)$ for integers a and b. Then, part W_{ni} is stored in the caches of all m ENs in subset $\mathcal{K}_{Ti}$.

In the delivery phase, consider a demand vector $\boldsymbol{d}$. With cooperative ZF precoding, each cluster can serve $u(m)$ users at a time. Based on this, the K_R users are grouped into

$$B_D = \frac{\text{l.c.m.}(u(m), K_R)}{u(m)} \tag{19}$$

groups, defined as $\{\mathcal{K}_{Rj}\}_{j=1}^{B_D}$[2], with

$$\mathcal{K}_{Rj} = \{[(j-1)u(m) + 1]_{K_R}, [(j-1)u(m) + 2]_{K_R}, \cdots, [ju(m)]_{K_R}\}. \tag{20}$$

Each cluster $\mathcal{K}_{Ti}$, $i \in [F_C]$ transmits the parts $\{W_{d_k i}\}_{k=1}^{K_R}$ of the requested files by serving each of the B_D groups of users in turn. To communicate to all $u(m)$ users in each group, each part $W_{d_k i}$ is further split equally into F_D packets as $W_{d_k i} = \{W_{d_k ij}\}_{j=1}^{F_D}$. With this split, each cluster of ENs shares $K_R F_D$ packets, which can be sent to B_D groups of users

[2]The set $\{\mathcal{K}_{Ti}\}_{i=1}^{F_C}$ is a 1-$(K_T, m, \text{l.c.m.}(u(m), K_R)/K_T)$ design, the set $\{\mathcal{K}_{Rj}\}_{j=1}^{B_D}$ is a 1-$(K_R, u(m), F_D)$ design (see [26]).

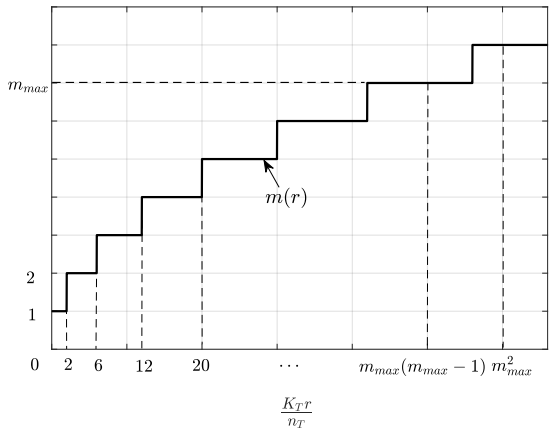

Fig. 3. Multiplicity $m(r)$ in (21) of the requested files obtained as a result of fronthaul transmission as a function of $K_T r/n_T$.

sequentially within $K_R F_D/u(m) = B_D$ blocks since $u(m)$ packets are sent in each block. The resulting number of blocks is $B = F_C B_D$, yielding the NDT in (7) $\delta_E = B/F = B_D/F_D$, as indicated in Proposition 1. ■

IV. Achievable NDT for Fronthaul-only caching

As a second preliminary result of interest, in this section, we present an achievable NDT (Proposition 2) for the case of no caching, i.e., with $\mu = 0$, as well as and the corresponding cloud-aided delivery scheme. We focus on serial fronthaul-edge transmission, while pipelined delivery will be considered in Section VI.

A. Achievable NDT

In the absence of caching, any desired multiplicity level $m \leq m_{max}$ can be realized for all the contents in the requested vector $\boldsymbol{d}$ thanks to fronthaul transmission. To this end, the fronthaul links are used to convey each packet of a requested file to a subset of m ENs. Choosing the subsets of ENs as described in the previous section (see (18)) allows the edge NDT (13) to be achieved thanks to cooperative EN transmission. Increasing m requires a larger fronthaul NDT since it requires to transfer more information on the fronthaul links, but it generally yields a lower edge NDT (13). The next proposition presents an achievable NDT obtained by optimizing over the values of m.

To elaborate, define the desired multiplicity for a given fronthaul rate r as

$$m(r) = \begin{cases} \left[\sqrt{\frac{K_T r}{n_T}}\right], & \text{for } r < r_{th} \\ m_{max}, & \text{for } r \geq r_{th}, \end{cases} \tag{21}$$

where $[x]$ represents the nearest positive integer function, and

$$r_{th} = \frac{n_T}{K_T} m_{max}^2. \tag{22}$$

The multiplicity (21) is illustrated in Fig 3. As seen, the selected multiplicity $m(r)$ is piece-wise constant and non-decreasing with respect to the fronthaul rate r. It is also respectively a non-decreasing an non-increasing function of K_T and n_T.

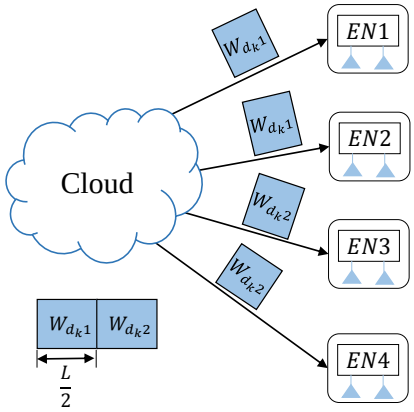

Fig. 4. Fronthaul transmission for the achievable NDT in Proposition 2 for $m(r) = 2$.

Proposition 2: For an F-RAN system with any fronthaul rate $r > 0$ and cache capacity $\mu = 0$, we have the upper bound on the minimum NDT

$$\delta^*(\mu = 0, r) \leq \delta_F(m(r)) + \delta_E(m(r)), \tag{23}$$

with the fronthaul NDT as a function of the multiplicity m given as

$$\delta_F(m) = \frac{K_R m}{K_T r}, \tag{24}$$

the multiplicity $m(r)$ in (21), and the edge NDT $\delta_E(m)$ defined in (13).

Proof: The proof is presented in Section IV-C. ■

B. Example

We now discuss an example that illustrates the proposed achievable cloud-aided delivery scheme. As in Example 1, we consider an F-RAN model with $K_T = 4$ ENs, $n_T = 2$ per-EN antennas, and $K_R = 4$ users, but with no caching, i.e., with $\mu = 0$ and $r > 0$.

Example 3. Consider $r = 2$, so we have the multiplicity $m(r) = 2$ from (21). Note that the multiplicity is the same as in Example 1 (see Fig. 2(a)). We hence use the same packetization and the same division of ENs into clusters of $m(r) = 2$ ENs discussed in Example 1, with the caveat that here only the packets of the requested files in $\boldsymbol{d}$ are made available to the ENs via fronthaul transmission. To elaborate, for any demand vector $\boldsymbol{d}$, each requested file W_{d_k} is divided into $F = 2$ equal and disjoint packets $\{W_{d_k1}, W_{d_k2}\}$ and the ENs are clustered into two groups, namely $\mathcal{K}_{T1} = \{1, 2\}$ and $\mathcal{K}_{T2} = \{3, 4\}$. With fronthaul transmission, packets $\{W_{d_k1}\}_{k=1}^{K_R}$ are sent to the ENs in cluster 1, and packets $\{W_{d_k2}\}_{k=1}^{K_R}$ to the ENs in cluster 2, as illustrated in Fig. 4. Hence, we have $\mathcal{F}_1 = \mathcal{F}_2 = \{W_{d_k1}\}_{k=1}^{K_R}$ and $\mathcal{F}_3 = \mathcal{F}_4 = \{W_{d_k2}\}_{k=1}^{K_R}$, and the resulting fronthaul NDT in (6) is $\delta_F = |\mathcal{F}_i|/(Fr) = 4/(2 \times 2) = 1$. For edge transmission, the cooperative delivery strategy in Example 1 can be applied by sending packets $\{W_{d_k1}\}_{k=1}^{4}$ and $\{W_{d_k2}\}_{k=1}^{4}$ sequentially in two blocks by the two clusters of ENs. Hence, the resulting edge NDT is $\delta_E = B/F = 2/2 = 1$, yielding the overall NDT $\delta_{ach}(\mu = 0, r) = \delta_F + \delta_E = 2$, as in (23).

C. Proof of Proposition 2

Fix a desired multiplicity m for the requested contents. Each requested file W_{d_k} is divided into F_C parts $W_{d_k} = \{W_{d_k i}\}_{i=1}^{F_C}$ with F_C in (15). Part $W_{d_k i}$ is sent to all the ENs in group $\mathcal{K}_{Ti}$ defined in (18) by using fronthaul transmission. Each part $W_{d_k i}$ is split into F_D equally sized packets packets $\{W_{d_k ij}\}_{j=1}^{F_D}$ with F_D in (16), and each EN i receives $|\mathcal{F}_i| = K_R F m/K_T$ packets with $F = F_C F_D$ in (17), so that the fronthaul NDT is $\delta_F(m) = |\mathcal{F}_i|/(Fr) = K_R m/(K_T r)$ as in (24). Transmission on the edge channels takes place as described in Section III-C, entailing the NDT $\delta_E(m)$ in (13). The overall NDT for a given multiplicity m is hence given as $\delta(m) = \delta_F(m) + \delta_E(m)$, which can be minimized over m. To this end, we define the function

$$\delta(x) = \frac{K_R x}{K_T r} + \frac{K_R}{x n_T}, \tag{25}$$

where $x \in [0, m_{max}]$ is a variable obtained by relaxing the integer constraints over m. The function $\delta(x)$ is convex within the range $[0, m_{max}]$, and the only stationary point is $x_0 = \sqrt{K_T r/n_T}$. Therefore, function $\delta(x)$ reaches its minimum at $x = x_0$. Based on this, the optimal multiplicity m is either $\lfloor x_0 \rfloor$ or $\lceil x_0 \rceil$ depends on whether $\delta(\lfloor x_0 \rfloor) < \delta(\lceil x_0 \rceil)$ or $\delta(\lfloor x_0 \rfloor) > \delta(\lceil x_0 \rceil)$, respectively. Hence, to simplify the analysis, the desired multiplicity m is chosen as the nearest positive integer of x_0, although this may not be optimal. This completes the proof. ■

V. Normalized delivery time Analysis on F-RAN

In Section III and IV, we have studied the special cases with edge caching only and no caching, respectively. In this section, we proceed to consider a general F-RAN model with any fronthaul rate $r \geq 0$ and cache capacity $\mu \geq 0$. We present an upper bound (Proposition 3) and a lower bound (Proposition 4) on the minimum NDT under serial delivery. These bounds provide a characterization of the minimum NDT that is conclusive for a wide range of values of the system parameters (Proposition 5) and is generally within a multiplicative factor of $3/2$ from optimality (Proposition 6). The main results offer insight into the optimal use of cloud and edge resources as a function of the fronthaul capacity, cache resources and number of ENs' transmit antennas.

A. Upper Bound on the Minimum NDT

In the presence of both cloud and edge resources, the multiplicity m of the requested files depends both on the pre-stored caching contents and on the information received at the ENs via fronthaul transmission. As a result, in an F-RAN, the optimal multiplicity is generally larger than the multiplicities $m(r)$ and $m(\mu)$ in (14) and (21), respectively. The multiplicity selected in the proposed scheme for any values of μ and r is given as

$$m(\mu, r) = \begin{cases} m(r), & \text{if } \mu K_T < m(r) \\ \lfloor \mu K_T \rfloor, & \text{if } m(r) \leq \mu K_T \leq m_{max} \\ m_{max}, & \text{if } \mu K_T > m_{max}. \end{cases} \tag{26}$$

The formula (26) has a graphical interpretation that will be discussed after the next proposition (see Fig. 5). Note that we have the equalities $m(\mu, 0) = m(\mu)$ and $m(0, r) = m(r)$.

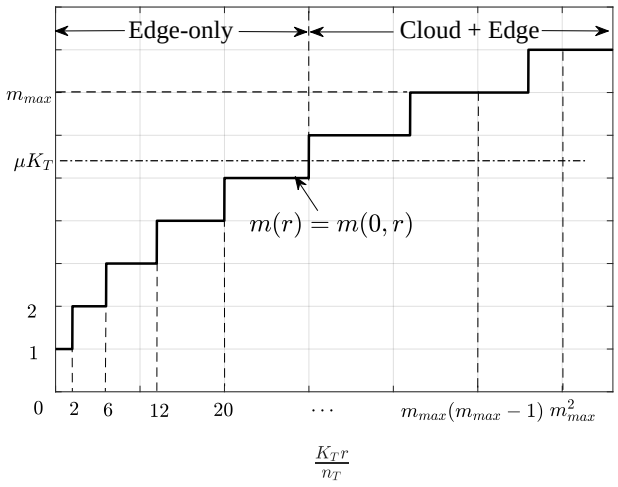

Fig. 5. Illustration of the multiplicity $m(\mu, r)$ (26) of the requested files, as a result of caching and fronthaul transmission, as selected by the proposed scheme: for $\mu K_T \leq m_{max}$, this is obtained by taking the maximum between $\lfloor \mu K_T \rfloor$ and $m(r)$ for the given value of r; while $\mu K_T > m_{max}$, we have $m(\mu, r) = m_{max}$.

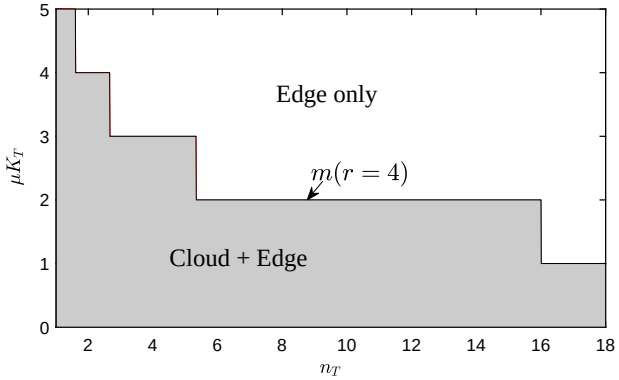

Fig. 6. The proposed caching-delivery scheme, which is approximately optimal by Proposition 3, leverages the cloud resources only for values of μ and n_T in the shaded area, while edge transmission is exclusively used otherwise ($K_T = 8, K_R = 40$ and $r = 4$).

Proposition 3: For an F-RAN system with any fronthaul rate $r \geq 0$ and cache capacity $\mu \geq 0$, we have the upper bound on the minimum NDT

$$\delta^*(\mu, r) \leq \delta_{ach}(\mu, r) = \text{l.c.e.}(\delta_F(\mu, r) + \delta_E(\mu, r)), \tag{27}$$

with the fronthaul NDT

$$\delta_F(\mu, r) = \delta_F(m(\mu, r) - \lfloor \mu K_T \rfloor), \tag{28}$$

with $\delta_F(m)$ in (24), and the edge NDT

$$\delta_E(\mu, r) = \delta_E(m(\mu, r)), \tag{29}$$

with $\delta_E(m)$ in (13).

Proof: The proof is presented in Appendix A. ■

According to (26), as illustrated in Fig. 5, the multiplicity $m(\mu, r)$ of each requested file is obtained by comparing μK_T, i.e., the multiplicity allowed by caching only, with the upper and lower bound m_{max} and the optimal multiplicity $m(r)$ selected

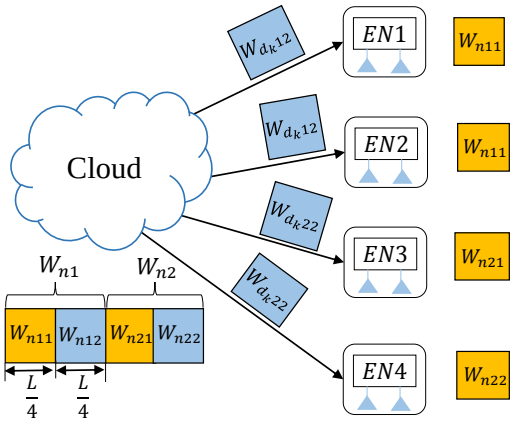

Fig. 7. Caching and delivery scheme under cloud and cache-aided transmission for the achievable NDT in Proposition 3 for $m(\mu, r) = m(r) = 3$.

when $\mu = 0$. We distinguish two cases.

- *Edge-only transmission:* For a sufficiently large cache capacity, i.e., when $\mu K_T \geq m(r)$, by (14) and (26), we have $m(\mu, r) = m(\mu)$, and hence the edge caches can support the selected multiplicity without the need for fronthaul transmission. In this case, we have $\delta_F(\mu, r) = 0$, and the NDT $\delta_{ach}(\mu, r)$ in (27) includes only the edge contribution $\delta_E(\mu, r)$;
- *Cloud and edge-aided transmission:* When $\mu K_T < m(r)$, we have the multiplicity $m(\mu, r) = m(r) > \mu K_T$, and hence fronthaul transmission is needed in order to support the multiplicity $m(\mu, r)$. In this case, the NDT $\delta_{ach}(\mu, r)$ in (27) includes both the contributions of fronthaul and edge NDTs.

Remark 1: From the discussion above, whenever the cache capacity is small enough to satisfy the inequality $\mu K_T < m(r)$, the proposed policy uses both cloud and edge resources. Accordingly, even when the edge alone would be sufficient to deliver all requested contents, that is, even when we have $\mu K_T \geq 1$, the policy uses cloud-to-edge communications if r is sufficiently large. This is because, in this regime, the cloud can send the uncached information to ENs in order to increase the multiplicity and hence to foster EN cooperation, at the cost of a fronthaul delay that does not offset the cooperation gains. However, when $\mu K_T \geq m(r)$, the scheme only uses edge resources. In fact, under this condition, the gains due to enhanced EN cooperation do not overcome the latency associated with fronthaul transmission. Fig. 6 illustrates the discussed conditions by depicting as shaded region of values of the pair (μ, n_T) for which inequality $\mu K_T < m(r)$ is satisfied and hence both cloud and edge transmission is used by the proposed scheme. An interesting observation is that, as n_T increases, edge processing becomes more effective, make cloud processing unnecessary for small values of the cache capacity μ. Note also that an increased r enlarges the region of values (μ, n_T) for which fronthaul transmission is used (not shown). ■

B. Example

Here we continue Example 1 and Example 3 by considering again an F-RAN with the example with $K_T = 4$ ENs, $n_T = 2$ per-EN antennas and $K_R = 4$ users, but in the general case with $\mu \geq 0$ and $r \geq 0$.

Example 4. Consider $\mu = 0.25$ and $r = 2$, so we have the multiplicity $m(\mu, r) = m(r) = 2$ by (26), which is the same as in both Example 1 and Example 3. To realize this multiplicity, we carry out first the same partition $\{W_{n1}, W_{n2}\}$ of each

library file W_n into two $F_C = 2$ parts as in Example 1 and Example 3. Moreover, here, each part is further split into $F_S = 2$ disjoint packets $\{W_{ni1}, W_{ni2}\}$ of equal size. In the placement phase, only packets $\{W_{ni1}\}_{n=1}^{N}$ for all the contents $\{W_n\}_{n=1}^{N}$ are cached at the ENs in cluster $\mathcal{K}_{Ti} = \{1, 2\}$ for $i = 1, 2$, as seen in Fig. 7. In the delivery phase, for any demand vector $\boldsymbol{d}$, the uncached packets $\{W_{d_k i2}\}_{k=1}^{K_R}$ of requested files are sent to the ENs in cluster $\mathcal{K}_{Ti}$. Therefore, each EN receives four packets on the fronthaul, yielding the fronthaul NDT $\delta_F = |\mathcal{F}_i|/(Fr) = 4/(4 \times 2) = 1/2$. Using clustered EN cooperation, packets $\{W_{d_k 1i}\}_{k=1}^{K_R}$ for $i = 1, 2$ can be sent by the ENs in cluster $\mathcal{K}_{T1}$ in two blocks, while packets $\{W_{d_k 2i}\}_{k=1}^{K_R}$ can be similarly delivered by the ENs in cluster $\mathcal{K}_{T2}$. As a result, the edge NDT is $\delta_E = B/F = 4/4 = 1$. Hence, the overall NDT is $\delta_{ach}(\mu, r) = \delta_F + \delta_E = 3/2$, as in (27).

C. Lower Bound on the Minimum NDT

A lower bound on the minimum $\delta^*(\mu, r)$ is presented in the following proposition, where we define the function $m^*(r)$ as

$$m^*(r) = \begin{cases} \max\left\{\sqrt{\frac{K_T r}{n_T}}, 1\right\}, & \text{for } r < r_{th} \\ m_{max}, & \text{for } r \geq r_{th}, \end{cases} \tag{30}$$

with r_{th} as in (22).

Proposition 4: In an F-RAN with n_T antennas at each transmitter, the minimum NDT $\delta^*(\mu, r)$ is lower bounded as

$$\delta^*(\mu, r) \geq \delta_{lb}(\mu, r) = \begin{cases} \max\left\{\frac{K_R(m^*(r) - \mu K_T)}{K_T r} + \frac{K_R}{m^*(r) n_T}, 1\right\}, & \text{for } \mu K_T < m^*(r) \quad (31a) \\ \max\left\{\frac{K_R}{\mu K_T n_T}, 1\right\}, & \text{for } \mu K_T \geq m^*(r). \quad (31b) \end{cases}$$

Proof: The proof is presented in Appendix B. ∎

D. Minimum NDT

The following proposition characterizes the minimum NDT $\delta^*(\mu, r)$ for the regime of low cache and fronthaul capacities, namely, when $\mu K_T \in [0, 1]$ and $r \in [0, n_T/K_R]$, as well as for any set-up with μK_T integer, or with sufficiently large caches such that $\mu K_T \geq m_{max}$.

Proposition 5: For an F-RAN system with n_T antennas at each EN, the minimum NDT $\delta^*(\mu, r)$ is given as

$$\delta^*(\mu, r) = \begin{cases} \max\left\{\frac{K_R(1 - \mu K_T)}{K_T r} + \frac{K_R}{n_T}, 1\right\}, & \text{for } \mu K_T \in [0, 1] \text{ and } r \in [0, \frac{n_T}{K_T}] \\ \max\left\{\frac{K_R}{\mu K_T n_T}, 1\right\}, & \text{for } \mu K_T \in \{m(r) + 1, \cdots, m_{max}\} \cup (m_{max}, K_T]. \end{cases} \tag{32}$$

Proof: The result follows by the direct comparison of the bounds in Proposition 1 and Proposition 2. ∎

More generally, the achievable NDT in Proposition 3 is within a factor of $3/2$ from minimum NDT for any fractional caching size μ and fronthaul rate r.

Proposition 6: For an F-RAN system with n_T antennas at each EN, and any value of $\mu \geq 0$ and $r \geq 0$, we have the inequality

$$\frac{\delta_{ach}(\mu, r)}{\delta^*(\mu, r)} \leq \frac{3}{2}. \tag{33}$$

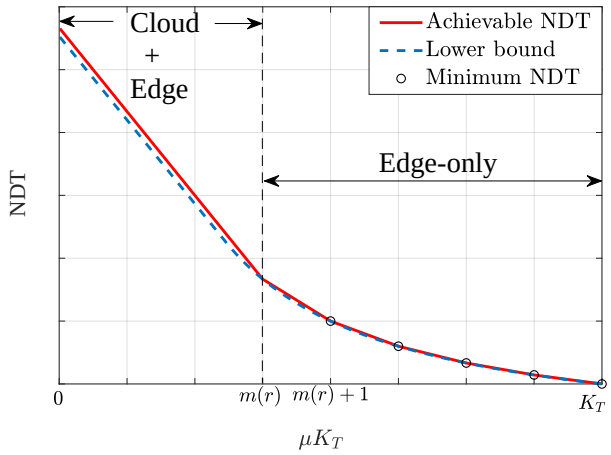

Fig. 8. Achievable NDT $\delta_{ach}(\mu, r)$ in Proposition 3 (solid curve), and lower bound on the minimum NDT $\delta^*(\mu, r)$ in Proposition 2 (dashed line) versus μ for a given value of r. The figure highlights the two regimes of values of the cache capacity μ with which the achievable schemes use edge-only or both cloud and edge transmission.

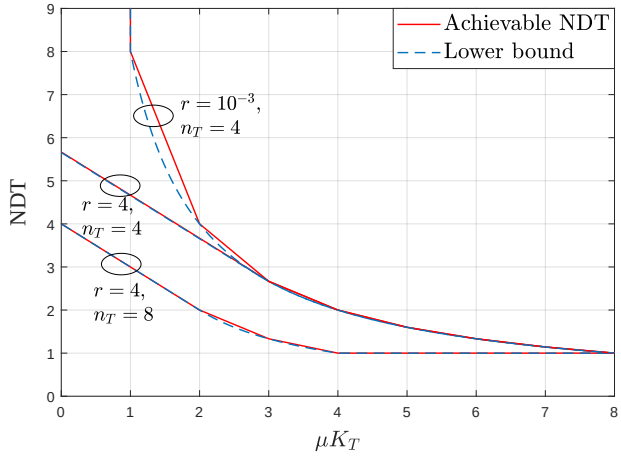

Fig. 9. Achievable NDT $\delta_{ach}(\mu, r)$ and lower bound $\delta_{lb}(\mu, r)$ versus μ for different values of r and n_T, with $K_T = 8$ and $K_R = 32$.

Proof: The proof is presented in Appendix D. ■

A plot of the achievable NDT $\delta_{ach}(\mu, r)$ and of the lower bound $\delta_{lb}(\mu, r)$ as a function of the fractional cache capacity μ is shown in Fig. 8 for a given value of r. As discussed, the achievable scheme uses both cloud and edge resources when $\mu K_T < m(r)$, while it uses only edge transmission when $\mu K_T \geq m(r)$. In the first regime, the fronthaul NDT decreases linearly with μK_T, which leads to a linear decrease in the overall NDT $\delta_{ach}(\mu, r)$. Instead, in the second regime, the achievable NDT $\delta_{ach}(\mu, r)$ is piece-wise linear and decreasing. For this range of values of μ, time-sharing between two successive multiplicities is carried out for delivery, unless μK_T is an integer. By comparison with the lower bound, the figure also highlights the regimes, identified in Proposition 5, in which the scheme is optimal.

The achievable NDT in Proposition 1 and lower bound in Proposition 2 are plotted in Fig. 9 as a function of μ for $K_T = 8$ and $K_R = 32$ and for different values of r and n_T. As stated in Proposition 3, the achievable NDT is optimal when μ and r are small enough, as well as when μ equals a multiple of $1/K_T = 1/8$ or is large enough. For values of r close to zero, the NDT diverges as μ tends to $1/K_T = 1/8$, since requests cannot be supported based solely on edge transmission. For larger values of r and/or n_T, the NDT decreases. In particular, when $\mu K_T \geq m_{max} = 4$ and $n_T = 8$, as discussed, we have the

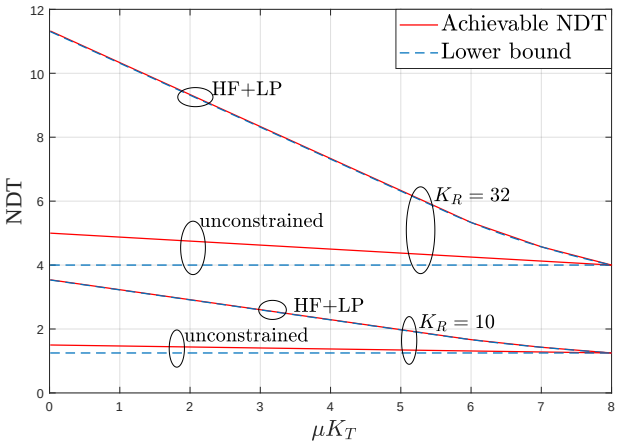

Fig. 10. Comparison of the bounds derived in this paper under the constraints of hard-transfer fronthauling (HF) and one-shot linear precoding (LP) with the bounds derived in [7] without such constraints, with $K_T = 8, n_T = 1$ and $r = 4$.

ideal NDT of one, since the maximum possible number 32 of users can be served.

Finally, for reference, a comparison of the bounds derived here under the assumptions of hard-transfer fronthauling (HF), i.e., the transmission of uncoded files on the fronthaul links, and of one-shot linear precoding (LP) with those derived in [7, Corollary 1 and Proposition 4] without such constraints, is illustrated in Fig. 10. We recall that the achievable scheme in [7] is based on fronthaul quantization and on delivery via interference alignment and ZF precoding. The figure is obtained for $K_T = 8, n_T = 1, r = 4$, and for different values of K_R. It is observed that the loss in performance caused by the practical constraints considered in this work is significant, and that it increases with the number K_R of users. This conclusion confirms the discussion in [7, Sec. IV-B].

VI. Pipelined fronthaul-edge transmission

In this section, we consider pipelined fronthaul-edge transmission, whereby fronthaul transmission from the cloud to the ENs and edge transmission from the ENs to the users can take place simultaneously. Note that this is possible due to the orthogonality of the two channels. We first describe how the system model and performance metric are modified as compared to the serial model of Section II, and then we present upper and lower bounds on the minimum NDT. As for the serial case, the bounds will reveal that the proposed cloud and cache-aided transmission policy is optimal for a large range of system parameters, and is generally within a multiplicative factor, here of two, from the information-theoretic optimal performance. Importantly, in contrast to the approximately optimal serial strategy, the proposed scheme for pipelined transmission leverages fronthaul transmission for any non-zero value of the fronthaul rate r and for any value of the fractional cache capacity μ less than one.

A. System Model and Performance Metric

The system model is as in Section II, with the main caveat that fronthaul and wireless transmissions can occur simultaneously. Specifically, caching is defined as in Section II by sets $\{\mathcal{F}_i\}_{i=1}^{K_T}$. As shown in Fig. 11, the overall delivery time for a given request vector is organized into B blocks. In any block $b \in [B]$, the cloud transmits the packets in set $\mathcal{F}_i(b) = \{\mathcal{F}_{id_1}(b), \cdots, \mathcal{F}_{id_K}(b)\}$

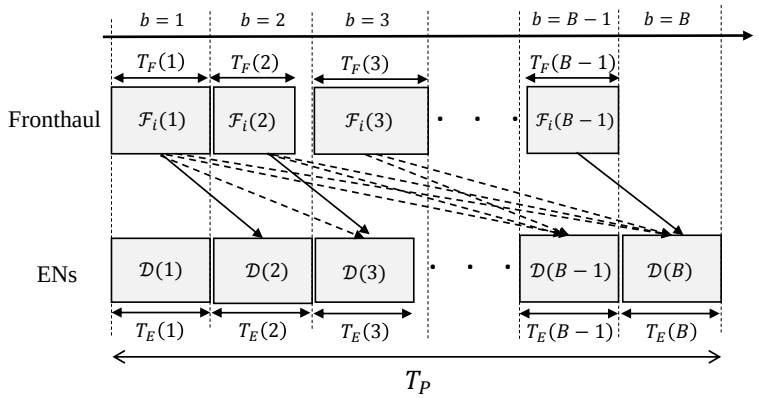

Fig. 11. Illustration of pipelined F-RAN operation for general strategies (both dashed and solid lines) and for block-Markov transmission (solid lines only).

to EN i, where $\mathcal{F}_{id_k}(b) \in \mathcal{F}_{id_k}$ is a subset of the packets requested by user k. In the same block, each EN i sends the subset $\mathcal{D}(b)$ of requested packets to a subset $\mathcal{R}(b)$ of users by utilizing the cached contents and fronthaul information $\{\mathcal{F}_i(b')\}_{i=1}^{K_T}$ received in the previous blocks $b' = 1, \ \cdots, \ b-1$. Note that the ENs can use the information received on the fronthaul in a causal way along the blocks. As in (4), the duration of the fronthaul transmission in each block b is given as

$$T_F(b) = \max_{i \in [K_T]} \frac{|\mathcal{F}_i(b)|L}{F} \frac{1}{C_F}, \tag{34}$$

and, following (5), the edge transmission time is given as the sum over the blocks

$$T_E(b) = \frac{L}{F} \frac{1}{(\log(P) + o(\log(P))}. \tag{35}$$

Since each block needs to accommodate both fronthaul and edge transmissions, the duration of a block is the maximum of the above two times, i.e., $T_P(b) = \max\{T_F(b), T_E(b)\}$. The total delivery time is hence given as

$$T_P = \sum_{b=1}^{B} T_P(b) = \sum_{b=1}^{B} \max\{T_E(b), T_F(b)\}. \tag{36}$$

Finally, following (6)-(7), the pipelined NDT δ_P is computed as the limit

$$\delta_P = \lim_{P \to \infty} \lim_{L \to \infty} \frac{T_P}{L/\log(P)}. \tag{37}$$

The minimum NDT $\delta_P^*(\mu, r)$ is defined as in (9)-(10). Following the same argument in [7, Lemma 4], the minimum NDT $\delta_P^*(\mu, r)$ pipelined delivery satisfies the inequalities $\delta^*(\mu, r)/2 \leq \delta_P^*(\mu, r) \leq \delta^*(\mu, r)$, and hence the improvement in NDT under pipelined transmission is upper bounded by a factor of two.

B. Achievable Scheme and Upper Bound on the Minimum NDT

In this section, we derive an achievable NDT by proposing a caching and delivery strategy that leverage simultaneous fronthaul and edge transmission. We start by observing that any serial strategy defined by the tuple $\{\mathcal{C}_i, \mathcal{F}_i, \{\boldsymbol{v}_{inf}(b)\}_{n \in [N], f \in [F], b \in [B]}\}_{i=1}^{K_T}$, as described in Section II, can be converted into a pipelined transmission strategy $\{\mathcal{C}_i, \{\mathcal{F}_i(b)\}_{b \in [B]}, \{\boldsymbol{v}_{inf}(b)\}_{n \in [N], f \in [F], b \in [B]}\}_{i=1}^{K_T}$. This is done by means of block-Markov transmission, as illustrated in

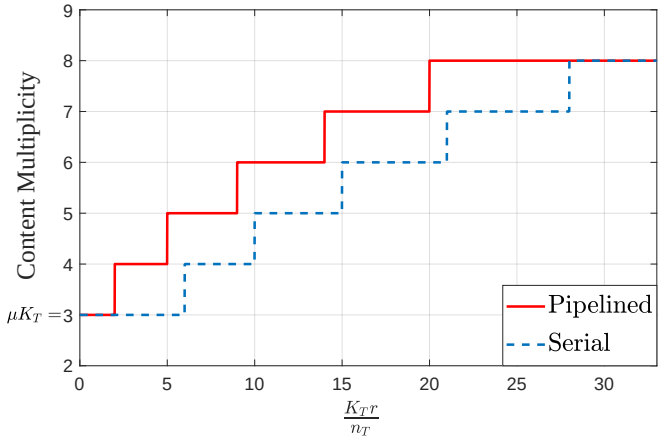

Fig. 12. Multiplicity $m_p(\mu, r)$ in (40) under pipelined transmission and $m(\mu, r)$ in (26) under serial transmissions for $K_T = 8, K_R = 32, n_T = 4$ and $\mu K_T = 3$.

Fig. 11. To this end, caching is carried out in the same way as in the serial strategy. For delivery, any packet in the fronthaul message $\mathcal{F}_i$ is split into $B-1$ equal subpackets, and each bth subpacket is placed in the set $\mathcal{F}_i(b)$ communicated in block $b = 1, 2, \cdots, B-1$. The edge delivery scheme for the serial strategy is applied in each following block $b = 2, \cdots, B$ to the subpackets received in the previous block $b-1$.

Suppose that the original serial scheme has fronthaul and edge NDTs δ_F and δ_E, respectively. Then, by the definitions (34) and (35), the contribution to the NDT for each block b is given as $\max\{\delta_F, \delta_E\}/(B-1)$, since each block delivers a fraction $1/(B-1)$ of each packet. Hence, the overall NDT is given as $\delta_P = (B/B-1)\max\{\delta_F, \delta_E\}$, which yields for $B \to \infty$ the NDT

$$\delta_P = \max\{\delta_F, \delta_E\}. \tag{38}$$

Based on the described block-Markov approach, as a first solution, one could convert the approximately optimal serial policy derived in the previous section to obtain an achievable NDT for the pipelined model. However, in the proposed serial strategy, it can be proved that the edge NDT $\delta_E(\mu, r)$ in (29) is generally larger than the fronthaul NDT $\delta_F(\mu, r)$ in (28). By (38), under pipelined transmission, the latency is determined by the maximum of fronthaul and edge latencies. Therefor, this first solution would leave open the possibility to increase the content multiplicity at the edge by sending more information through the fronthaul links to the ENs without increasing the system latency. This observation is leveraged by the proposed scheme.

To start, we define a multiplicity $m_p(\mu, r)$ for the pipelined model, which is no less than the serial multiplicity $m(\mu, r)$ in (26). This is done by first evaluating the multiplicity that ensures that fronthaul and edge transmission NDTs in (28) and (29), respectively, are equal, obtaining

$$m_{eq}(\mu, r) = \frac{\mu K_T}{2} + \frac{\sqrt{(\mu K_T n_T)^2 + 4 n_T K_T r}}{2 n_T}. \tag{39}$$

In order to account for the maximum multiplicity m_{max} (12) and for the requirement that the multiplicity be an integer, we

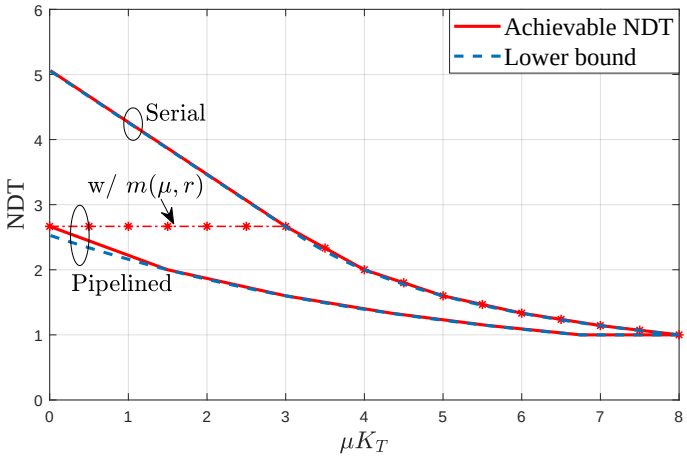

Fig. 13. Achievable NDTs under serial and pipelined transmissions with different multiplicity choices for $K_T = 8, K_R = 32, n_T = 4$ and $r = 5$.

then define the multiplicity adopted by the proposed scheme as

$$m_p(\mu, r) = \begin{cases} \max\{\lfloor m_{eq}(\mu, r) \rfloor, 1\}, & \text{for } \mu K_T \leq m_{max} - \frac{K_T r}{m_{max} n_T} \\ m_{max}, & \text{for } \mu K_T \geq m_{max} - \frac{K_T r}{m_{max} n_T}. \end{cases} \tag{40}$$

As an example, the multiplicities $m(\mu, r)$ in (26) and $m_p(\mu, r)$ in (40) under serial and pipelined transmissions, respectively, are plotted in Fig. 12 for $K_T = 8, K_R = 32, n_T = 4$ and $\mu K_T = 3$. Both multiplicities increase with r from the multiplicity μK_T that can be ensured by the edge cache resource only. The figure confirms that the proposed pipelined transmission scheme increases the multiplicity by leveraging simultaneous fronthaul and edge transmissions. The resulting achievable NDT is presented in the following proposition.

Proposition 7: For an F-RAN system with n_T antennas, we have the upper bound $\delta_p^*(\mu, r) \leq \delta_{p,ach}(\mu, r)$ on the minimum NDT for pipelined fronthaul-edge transmission, where

$$\delta_{p,ach}(\mu, r) = \begin{cases} \max\left\{\frac{K_R(1-\mu K_T)}{K_T r}, 1\right\}, & \text{for } \mu K_T \leq \left(1 - \frac{K_T r}{n_T}\right)^+ \\ \text{l.c.e.}\left\{\max\left\{\frac{K_R}{m_p(\mu, r) n_T}, 1\right\}\right\}, & \text{for } \mu K_T \geq \left(1 - \frac{K_T r}{n_T}\right)^+ \end{cases} \tag{41}$$

Proof: As discussed, the proposed scheme leverages block-Markov delivery and uses the multiplicity (40). As a result, the NDT is given by (38) with δ_F in (28) and δ_E in (29), where the multiplicity is in (40). Note that, with this scheme, for the small cache regime of $\mu K_T \leq (1 - K_T r/n_T)^+$, by (40), we have the multiplicity $m_p(\mu, r) = 1$ for each block, and the system performance is dominated by the fronthaul NDT δ_F in (28). Instead, for $\mu K_T \geq (1 - K_T r/n_T)^+$, when the multiplicity $m_p(\mu, r)$ is an integer, i.e., when $m_p(\mu, r) = m_{eq}(\mu, r)$, the fronthaul and edge NDTs in (28) and (29) are equal. When $m_{eq}(\mu, r)$ is not an integer, time sharing is performed between the two integer multiplicities $\lfloor m_{eq}(\mu, r) \rfloor$ and $\lceil m_{eq}(\mu, r) \rceil$. ■

A comparison of the achievable NDTs under serial and pipelined transmissions for the same system parameters as in Fig. 12 can be found in Fig. 13. Beside the achievable NDTs in (27) and (41), we also plot for reference the NDT of the pipelined policy obtained by using the same multiplicity $m(\mu, r)$ in (26) of the serial policy. Pipelining is seen to bring a non-negative reduction in NDT as compared to the serial policy even when the multiplicity is not optimized due to the possibility to use fronthaul and edge transmissions simultaneously. However, as discussed, the NDT performance with multiplicity $m(\mu, r)$ is

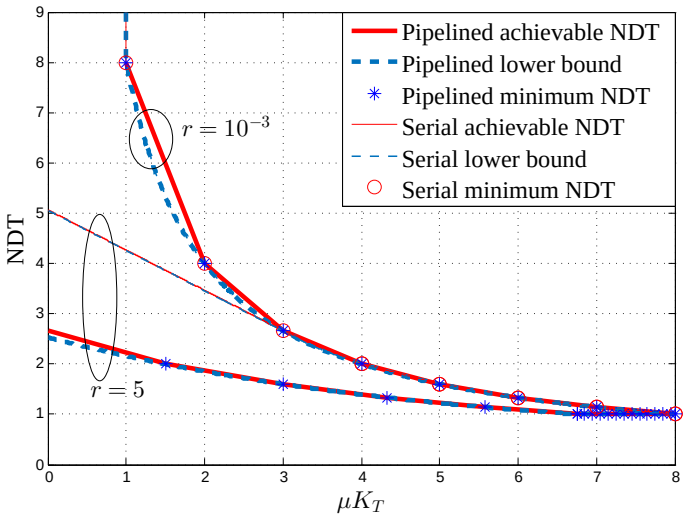

Fig. 14. Achievable NDT $\delta_{p,ach}(\mu, r)$ in Proposition 7 and lower bound $\delta_{p,lb}(\mu, r)$ Proposition 8 versus μ for different values of r, with $K_T = 8, K_R = 32$, and $n_T = 4$.

dominated by edge transmission. As a result, when $\mu K_T \geq m(r) = 3$, this policy does not provide NDT gains since, in this regime, only edge resources are used, i.e., $\delta_F = 0$. In contrast, the proposed policy, with the multiplicity $m_p(\mu, r)$ in Fig. 12 can improve NDT for all values of $\mu \in [0, 1]$.

Pipelined and serial NDTs (41) and (27) are further compared in Fig. 14 as a function of μ for two different values of r for $K_T = 8, K_R = 32$ and $n_T = 4$. For small values of r, such as $r = 10^{-3}$ in the figure, the benefits from pipelining transmission is limited due to the bottleneck posed by the low fronthaul rate. Instead, for larger values of r, the reduction in NDT is evident. In particular, with pipelined delivery, when $\mu K_T \geq m_{max} - K_T r/(m_{max} n_T) = 6.75$, the maximum multiplicity m_{max} can be supported, yielding the ideal NDT of one. In contrast, for the serial case, the ideal NDT of one can be achieved only with full caching, i.e., $\mu = 1$. Finally, as μK_T increases, edge transmission becomes more efficient, reducing the contribution from fronthaul transmission and yielding a smaller improvement due to pipelined transmission.

Remark 2: The proposed scheme is able to provide the discussed latency gains by leveraging *cloud and edge-aided transmission* for any value of μ and r except for the extreme case $\mu K_T \geq m_{max}$, where the maximum multiplicity m_{max} can be ensured by caching only. We emphasize that this stands in stark contrast to the serial case in which, as observed in Remark 1, when $\mu K_T \geq m(r)$, the overhead due to fronthaul transmission becomes excessive and edge transmission alone is preferable.

C. Lower Bound on the Minimum NDT

A lower bound on the minimum NDT $\delta_p^*(\mu, r)$ is presented in the following proposition, where we define the function $m_p^*(\mu, r)$ as

$$m_p^*(\mu, r) = \begin{cases} \max\{m_{eq}(\mu, r), 1\}, & \text{for } \mu K_T \leq m_{max} - \frac{K_T r}{m_{max} n_T} \\ m_{max}, & \text{for } \mu K_T \geq m_{max} - \frac{K_T r}{m_{max} n_T}, \end{cases} \tag{42}$$

with $m_{eq}(\mu, r)$ as in (39).

Proposition 8: For an F-RAN system with n_T antennas, the minimum NDT $\delta_p^*(\mu, r)$ for pipelined fronthaul-edge transmission is lower bounded as

$$\delta_p^*(\mu,r) \geq \delta_{p,lb}(\mu,r) = \begin{cases} \max\{\frac{K_R(1-\mu K_T)}{K_T r}, 1\}, & \text{for } \mu K_T \leq (1-\frac{K_T r}{n_T})^+ \\ \max\left\{\frac{K_R}{m_p^*(\mu,r) n_T}, 1\right\}, & \text{for } \mu K_T \geq (1-\frac{K_T r}{n_T})^+. \end{cases} \tag{43}$$

Proof: The proof is presented in Appendix E. ■

D. Minimum NDT

The minimum NDT is characterized in the following Proposition for the regime of low cache and fronthaul capacities, i.e., with $\mu K_T \leq 1 - K_T r/n_T$, and for any set-up with $\mu K_T = i - K_T r/(i n_T)$, $i \in [m_{max}]$ or with large caches, i.e., with $\mu K_T \geq m_{max} - K_T r/(m_{max} n_T)$. These conditions are akin to those identified in Proposition 5 for the serial case, with the different definitions being due to the use of fronthaul resources for all system parameters.

Proposition 9: For an F-RAN system with N_T antennas, the minimum NDT $\delta_p^*(\mu, r)$ for pipelined fronthaul-edge transmission is given as

$$\delta_p^*(\mu,r) = \begin{cases} \max\left\{\frac{K_R(1-\mu K_T)}{K_T r}, 1\right\}, & \text{for } \mu K_T \leq 1 - \frac{K_T r}{n_T} \\ \max\left\{\frac{K_R}{m_p^*(\mu,r) n_T}, 1\right\}, & \text{for } \mu K_T \in \{i - \frac{K_T r}{i n_T}\}_{i=1}^{m_{max}} \cup (m_{max} - \frac{K_T r}{m_{max} n_T}, K_T]. \end{cases} \tag{44}$$

Proof: The result can be obtained by the direct comparison of the bounds in Proposition 7 and Proposition 8. ■

The optimality regimes are illustrated in Fig. 14. We emphasize that, in a manner similar to the serial case, for integer values of the multiplicity $m_p(\mu, r) = i$, the optimal NDT is achieved. The difference is that this multiplicity here is obtained with a cache size $\mu K_T = i - K_T r/(i n_T)$, where $K_T r/(i n_T)$ is the contribution to the multiplicity due to fronthaul transmission.

The general multiplicative gap between the performance of the proposed scheme and the minimum NDT is stated in the following proposition.

Proposition 10: For a general F-RAN system with n_T antennas at each EN, and any value of $\mu \geq 0$ and $r \geq 0$, we have the inequality

$$\frac{\delta_{p,ach}(\mu,r)}{\delta_p^*(\mu,r)} \leq 2. \tag{45}$$

Proof: The proof is presented in Appendix F. ■

VII. Conclusions

In fog-aided cellular systems, fronthaul resources enable a cloud processor with access to the content library to communicate uncached contents to the edge nodes. This information is not only necessary to enable content delivery when the overall system's capacity is insufficient, but it can also facilitate cooperative interference management. In this paper, we have studied the resulting optimal trade-off between fronthaul latency overhead and overall delivery latency from an information-theoretic viewpoint under the assumption of multi-antenna edge nodes, uncoded caching and fronthaul and one-shot linear precoding on the wireless edge channel. The minimum delivery latency was investigated in the high-SNR regime under both serial and

pipelined transmission models. The main results of this F-RAN model are the characterizations, within small multiplicative factors, of the minimum high-SNR latency as a function of system parameters such as fronthaul capacity, edge cache capacity and number of per-edge node antennas. Extensive numerical results have been provided to demonstrate the usefulness of the derived information-theoretic characterizations in understanding the interplay and relative roles of edge and cloud resources on the performance of fog-aided networks. We have also commented on the impact of the practical assumptions made here as compared to the unconstrained delivery strategies studied in [7].

The information-theoretic characterizations derived in this work leave open a number of research questions. A first line of work that has recently been partially addressed in [17], [18], [27] concerns the effect of imperfect or no CSI on the optimal design of caching and delivery techniques. A second, related, issue is the study of optimal edge caching techniques under partial connectivity [27], [28]. Third, the analysis can be extended to yield insights into the performance of online caching strategies by following [25]. Lastly, using ideas from [5], it would be interesting to generalize the results of this work to a set-up that includes also caching at the users.

APPENDIX A
PROOF OF PROPOSITION 3

As discussed in Section V, for a desired multiplicity $m \geq \lfloor \mu K_T \rfloor$, we distinguish the cases $m = \lfloor \mu K_T \rfloor$ and $m > \lfloor \mu K_T \rfloor$. In the first case, edge-only transmission is used, and we adopt the same cache-aided delivery strategy described in Section III-C, yielding the edge NDT in (13).

In contrast, for the case $m > \lfloor \mu K_T \rfloor$, cloud and edge-aided transmission is used, and the multiplicity m is obtained using both caching and fronthaul transmission during the delivery phase. Generalizing Example 4, we first divide each content into F_C parts $\{W_{ni}\}_{i=1}^{F_C}$, with F_C in (15), and then we further divide each part into

$$F_S = \text{l.c.m.}(F_D, m) \tag{46}$$

packets $\{W_{nij}\}_{j=1}^{F_S}$, with F_D defined in (16). As a result, the overall number of packets is

$$F = F_C F_S. \tag{47}$$

By (47), we have the inequalities $K_T \leq F \leq K_T K_R m$.

In the caching phase, the F_S packets are arbitrarily divided into two disjoint subsets W_{ni}^1 and W_{ni}^2, where subset W_{ni}^1 contains an integer number $F_S \lfloor \mu K_T \rfloor / m$ of packets and subset W_{ni}^2 contains the rest. During the caching phase, all the packets in subsets $\{W_{ni}^1\}_{n=1}^N$ are cached at all m EN in cluster $\mathcal{K}_{Ti}$ by following (18), while the packets in subsets $\{W_{ni}^2\}_{n\in[N],i\in[F_C]}$ are left uncached.

During the delivery phase, for a demand vector $\boldsymbol{d}$, the uncached $K_R F_S(1 - \lfloor \mu K_T \rfloor / m)$ packets in subsets $\{W_{d_k i}^2\}_{k\in[K_R]}$, with $i \in [F_C]$, are sent to all m ENs in cluster $\mathcal{K}_{Ti}$ as in (18). Hence, each EN receives $K_R F(m - \lfloor \mu K_T \rfloor)/K_T$ packets on the fronthaul, yielding the fronthaul NDT $\delta_F(m) = |\mathcal{F}_i|/Fr = K_R(m - \lfloor \mu K_T \rfloor)/(K_T r)$. As a result of fronthaul transmission, for each file $W_{d_k}, d_k \in \boldsymbol{d}$, the ENs in each cluster $\mathcal{K}_{Ti}$ share all F_S packets in subsets $\{W_{d_k i}^1\}$ and $\{W_{d_k i}^2\}$. These ENs can

hence transmit cooperatively to the B_D groups $\{\mathcal{K}_{Rj}\}_{j=1}^{B_D}$ of $u(m) = mn_T$ users defined in (20) by using $(F_S/F_D)B_D$ blocks. Hence, the total number of blocks is

$$B = \frac{F_S}{F_D} B_D F_C, \tag{48}$$

yielding the edge NDT in (7), i.e., $\delta_E(m) = B/F = B_D/F_D = K_R/(mn_T)$.

As a result, for a given multiplicity m, the overall NDT is given as $\delta(m) = \delta_E(m)+\delta_F(m)$ with integer $m \in [\lfloor \mu K_T \rfloor, m_{max}]$. To minimize the NDT $\delta(m)$, we define the function $\delta(x)$ as

$$\delta(x) = \frac{K_R(x - \lfloor \mu K_T \rfloor)}{K_T r} + \frac{K_R}{x n_T}, \tag{49}$$

where $x \in [\lfloor \mu K_T \rfloor, m_{max}]$ is a variable obtained by relaxing the integer constraints over m. Function $\delta(x)$ is convex within its domain, and has only one stationary point $x_0 = \sqrt{K_T r/n_T}$. Hence, it reaches the minimum at point $x = x_0$ for $\lfloor \mu K_T \rfloor \leq x_0$, and point $x = \lfloor \mu K_T \rfloor$ for $\lfloor \mu K_T \rfloor \geq x_0$. While in the latter case the solution is an integer, in the former case, the optimal solution is given as $\lfloor x_0 \rfloor$ if $\delta(\lfloor x_0 \rfloor) < \delta(\lceil x_0 \rceil)$ or $\lceil x_0 \rceil$ if $\delta(\lfloor x_0 \rfloor) > \delta(\lceil x_0 \rceil)$. Here, in order to simplify the expressions, we select $[x_0]$ when $\lfloor \mu K_T \rfloor \leq m(r)$.

APPENDIX B
PROOF FOR PROPOSITION 2

The proof follows [5, Section 5] with the important caveats that here we need to additionally consider the delivery latency due to fronthaul transmission, as well as the extension to the general case $n_T \geq 1$. To start, we consider an arbitrary split of each file into $2^{K_T} - 1$ parts, such that each part $W_{n\tau}$, indexed by a subset $\tau \subseteq [K_T]$, contains an integer number of packets, including possibly no packets. We recall that each packet contains L/F bits. Part $W_{n\tau}$ is available at the ENs in the subset τ, either from the edge caches or from the cloud after fronthaul transmission. Note that this partition comes with no loss of generality, since each packet W_{nf} is available at all EN i such that $W_{nf} \in \mathcal{C}_i \cup \mathcal{F}_i$ (see definitions in Section II-A).

To distinguish between the contributions of cache and fronthaul resources, we use $c_{n\tau}$ to denote the number of cached packets from file W_n at the ENs in subset τ; while $f_{n\tau}(\boldsymbol{d})$ is the number of packets of file W_n sent on the fronthaul links of all ENs in subset τ for a given demand vector $\boldsymbol{d}$. Hence, part $W_{n\tau}$ has $a_{n\tau} = c_{n\tau} + f_{n\tau}(\boldsymbol{d})$ packets in total. The variables $\{c_{n\tau}\}$ and $\{f_{n\tau}(\boldsymbol{d})\}$, for all $n \in [N], \tau \subseteq [K_T]$ and vectors $\boldsymbol{d}$, fully specify the operation of the cache strategy $\mathcal{C}_i$ and fronthaul policy $\mathcal{F}_i$ defined in Section II-A.

Minimizing the NDT with respect to the caching strategy $\{c_{n\tau}\}_{n\in[N],\tau\subseteq\mathcal{T}}$ and fronthaul policy $\{f_{n\tau}(\boldsymbol{d})\}_{n\in\boldsymbol{d},\tau\subseteq\mathcal{T}}$ for all

vectors $\boldsymbol{d}$ yields the following integer problem

$$\underset{\{c_{n\tau}\},\{f_{n\tau}(\boldsymbol{d})\}}{\text{minimize}} \max_{\boldsymbol{d}} \ \delta_E^*\big(\{c_{n\tau}\},\{f_{n\tau}(\boldsymbol{d})\},\boldsymbol{d}\big)+\delta_F^*(\boldsymbol{d}) \tag{50a}$$

$$\text{s.t.} \sum_{i=1}^{K_T} \sum_{\substack{\tau\subseteq[K_T]:\\ |\tau|=i}} (c_{n\tau}+f_{n\tau}(\boldsymbol{d}))=F, \forall n\in\boldsymbol{d}, \forall \boldsymbol{d} \tag{50b}$$

$$\sum_{n=1}^{N} \sum_{\substack{\tau\subseteq[K_T]:\\ i\in\tau}} c_{n\tau} \le \mu F N, \forall i\in[K_T] \tag{50c}$$

$$\frac{1}{Fr}\sum_{n\in\boldsymbol{d}} \sum_{\substack{\tau\subseteq[K_T]:\\ i\in\tau}} f_{n\tau}(\boldsymbol{d}) \le \delta_F^*(\boldsymbol{d}), \forall i\in[K_T], \forall \boldsymbol{d} \tag{50d}$$

$$c_{n\tau}\ge 0, f_{n\tau}(\boldsymbol{d})\ge 0 \tag{50e}$$

$$0\le \delta_F^*(\boldsymbol{d}) \le \delta_{Fmax}, \tag{50f}$$

where $\delta_E^*\big(\{c_{n\tau}\},\{f_{n\tau}(\boldsymbol{d})\},\boldsymbol{d}\big)$ is the minimum edge NDT (7) for given cache and fronthaul policies when the request vector is $\boldsymbol{d}$. In (50b), the equality constraints enforce that all F packets of each requested file are available collectively at the ENs after the fronthaul transmission; inequalities (50c) come from the fact that the size of the cache content $\mathcal{C}_i$ of each EN i, which is given as $\sum_{n=1}^{N}\sum_{\tau\subseteq[K_T]:i\in\tau} c_{n\tau}$, is constrained by the cache capacity $\mu F N$ (see (1)); inequalities (50d) follow from the definition of fronthaul NDT (6), since the left-hand side is the number of packets sent to EN i on the fronthaul for request vector $\boldsymbol{d}$; and inequalities (50f) impose that the fronthaul NDT is no larger than

$$\delta_{Fmax}=\frac{K_R(m_{max}-\mu K_T)^+}{K_T r}. \tag{51}$$

This is because, as discussed in Section V-A, the multiplicity of the requested files can be upper bounded without loss of generality by m_{max}, and the maximum overall number of bits that are needed from the cloud to ensure this multiplicity is given as $K_R(m_{max}-\mu K_T)^+ L$ bits.

The optimum value of optimization problem (50) is lower bounded by substituting the maximum over all the request vector $\boldsymbol{d}$ with an average. In particular, since the number of ways to request all the K_R distinct files out of N library files is $\pi(N,K_R)=N!/(N-K_R)!$, the lower-bounding problem can be written as

$$\underset{\{c_{n\tau}\},\{f_{n\tau}(\boldsymbol{d})\}}{\text{minimize}} \frac{1}{\pi(N,K_R)}\sum_{\boldsymbol{d}} \delta_E^*\big(\{c_{n\tau}\},\{f_{n\tau}(\boldsymbol{d})\},\boldsymbol{d}\big)+\delta_F^*(\boldsymbol{d}) \tag{52a}$$

$$\text{s.t. (50b)} - \text{(50f)}. \tag{52b}$$

We now obtain a lower bound on the optimal value of problem (52) and hence also of problem (50). To this end, we first bound the minimum edge NDT $\delta_E^*\big(\{c_{n\tau}\},\{f_{n\tau}(\boldsymbol{d})\},\boldsymbol{d}\big)$ in (52a) by studying the number of packets that can be served in each block as a function of the availability of files at the ENs.

Lemma 1: Consider a single edge transmission block b in which a set $\{W_{n_l f_l}\}_{l=1}^{L}$ of L packets are sent to L distinct users in set $\mathcal{R}(b)\subseteq[K_R]$. In order for each user in $\mathcal{R}(b)$ to be able to decode the desired packet without interference at the end of

the block, the number L of packets must be upper bounded as

$$L \leq \min_{l\in[L]} |\tau_l| n_T, \tag{53}$$

where for any packet $W_{n_l f_l}$, τ_l denotes the subset of ENs that have access to it, either as part of the pre-stored contents at the EN's cache or of the fronthaul received signals, i.e., $W_{n_l f_l} \in \{\mathcal{C}_i \cup \mathcal{F}_i\}_{i\in\tau_l}$.

Proof: The proof follows from [5, Lemma 3] with the following differences. For a block b, each EN i sends

$$\boldsymbol{x}_i(b) = \sum_{l:i\in\tau_l} \boldsymbol{v}_{in_l f_l}(b) s_{n_l f_l}(b), \tag{54}$$

and the received signal at user $k \in \mathcal{R}(b)$, is given as

$$y_k(b) = \sum_{i=1}^{K_T} \boldsymbol{h}_{ki}^T(b)\boldsymbol{x}_i(b) + z_k(b) \tag{55}$$

$$= \sum_{i=1}^{K_T} \boldsymbol{h}_{ki}^T(b) \sum_{l:i\in\tau_l} \boldsymbol{v}_{in_l f_l}(b) s_{n_l f_l}(b) + z_k(b) \tag{56}$$

$$= \sum_{l=1}^{L} \sum_{i\in\tau_l} \boldsymbol{h}_{ki}^T(b) \boldsymbol{v}_{in_l f_l}(b) s_{n_l f_l}(b) + z_k(b). \tag{57}$$

From (57), the channel can be considered as a multi-antenna broadcast channel with L transmitters, each having $|\tau_l| n_T$ antennas, that are connected to L single-antenna users. By following the same steps as in [5, Eq. (28)-(36)] the proof is completed. ■

Each subset τ of ENs needs to deliver parts $\{W_{n,\tau}\}$, $n \in \boldsymbol{d}$, which consists of a total of $\sum_j (c_{d_j\tau} + f_{d_j\tau})$ packets. From Lemma 1, the number of necessary blocks is at least $\sum_j (c_{d_j\tau} + f_{d_j\tau})/(|\tau| n_T)$. By summing over all subsets τ and applying (7), the minimum edge NDT $\delta_E^*(\{c_{n\tau}\}, \{f_{n\tau}(\boldsymbol{d})\}, \boldsymbol{d})$ can be lower bounded as

$$\delta_E^*\big(\{c_{n\tau}\}, \{f_{n\tau}(\boldsymbol{d})\}, \boldsymbol{d}\big) \geq \frac{1}{F} \sum_{i=1}^{K_T} \sum_{j=1}^{K_R} \sum_{\substack{\tau\subseteq[K_T]:\\ |\tau|=i}} \frac{c_{d_j,\tau} + f_{d_j,\tau}}{i n_T}. \tag{58}$$

This bound is instrumental in proving the following lemma, which completes the proof upon combination with the trivial lower bound $K_R/\min\{K_T n_T, K_R\}$ on the edge NDT.

Lemma 2: The optimal value of the problem (52) is lower bounded by

$$f_{min} = \begin{cases} \frac{K_R(m^*(r)-\mu K_T)}{K_T r} + \frac{K_R}{m^*(r) n_T}, & \mu K_T < m^*(r) \\ \frac{K_R}{\mu K_T n_T}, & \mu K_T \geq m^*(r), \end{cases} \tag{59}$$

where $m^*(r)$ is defined in (30).

Proof: The proof is presented in Appendix C. ■

APPENDIX C
PROOF OF LEMMA 2

We lower bound the two terms in (52a) separately by starting with the minimum average edge NDT

$$\frac{1}{\pi(N,K_R)}\sum_{\boldsymbol{d}}\delta_E^*\big(\{c_{n\tau}\},\{f_{n\tau}(\boldsymbol{d})\},\boldsymbol{d}\big) \tag{60a}$$

$$\overset{(a)}{\geq}\frac{1}{F\pi(N,K_R)}\sum_{i=1}^{K_T}\frac{1}{in_T}\left[\sum_{\boldsymbol{d}}\sum_{j=1}^{K_R}\sum_{\substack{\tau\subseteq[K_T]:\\|\tau|=i}}\big(c_{d_j\tau}+f_{d_j\tau}(\boldsymbol{d})\big)\right] \tag{60b}$$

$$\overset{(b)}{=}\frac{1}{F\pi(N,K_R)}\sum_{i=1}^{K_T}\frac{1}{in_T}\left[K_R\sum_{\substack{\tau\subseteq[K_T]:\\|\tau|=i}}\pi(N-1,K_R-1)\sum_{n=1}^{N}(c_{n\tau}+\tilde{f}_{n\tau})\right] \tag{60c}$$

$$=\frac{K_R}{NF}\sum_{i=1}^{K_T}\frac{1}{in_T}\left[\sum_{\substack{\tau\subseteq[K_T]:\\|\tau|=i}}\sum_{n=1}^{N}(c_{n\tau}+\tilde{f}_{n\tau})\right] \tag{60d}$$

$$\overset{(c)}{=}\frac{K_R}{NFn_T}\sum_{i=1}^{K_T}\frac{1}{i}b_i \tag{60e}$$

$$\overset{(d)}{\geq}\frac{K_R}{NFn_T}\frac{(\sum_{i=1}^{K_T}b_i)^2}{\sum_{i=1}^{K_T}ib_i}, \tag{60f}$$

where inequality (a) follows from inequality (58); equality (b) holds because, for any library file W_n, the number of different request vectors that include file W_n is $K_R\pi(N-1,K_R-1)$, i.e., $\sum_{\boldsymbol{d}}\sum_{j=1}^{K_R}W_{d_j\tau}=K_R\pi(N-1,K_R-1)\sum_{n=1}^{N}W_{n\tau}$, and hence we have

$$\sum_{\boldsymbol{d}}\sum_{j=1}^{K_R}c_{d_j\tau}=K_R\pi(N-1,K_R-1)\sum_{n=1}^{N}c_{n\tau} \tag{61a}$$

$$\text{and}\quad\sum_{\boldsymbol{d}}\sum_{j=1}^{K_R}f_{d_j\tau}(\boldsymbol{d})=\sum_{\boldsymbol{d}}\sum_{n\in\boldsymbol{d}}f_{n\tau}(\boldsymbol{d})=K_R\pi(N-1,K_R-1)\sum_{n=1}^{N}\tilde{f}_{n\tau}, \tag{61b}$$

where $\tilde{f}_{n\tau}=\sum_{\boldsymbol{d}:n\in\boldsymbol{d}}f_{n\tau}(\boldsymbol{d})/(K_R\pi(N-1,K_R-1))$ represents the number of packets in part $W_{n\tau}$ for each user in τ received from the cloud; equality (c) follows the definition

$$b_i=\sum_{\substack{\tau\subseteq[K_T]:\\|\tau|=i}}\sum_{n=1}^{N}(c_{n\tau}+\tilde{f}_{n\tau}); \tag{62}$$

and inequality (d) applies the Cauchy-Schwarz inequality $(\sum_{i=1}^{n}u_iv_i)^2\leq(\sum_{i=1}^{n}u_i^2)(\sum_{i=1}^{n}v_i^2)$ by setting $u_i=\sqrt{b_i/i}$ and $v_i=\sqrt{ib_i}$.

To compute the term $\sum_{i=1}^{K_T} b_i$ in (60f), we impose the constraint (50b), obtaining

$$\pi(N,K_R)K_R F \overset{(a)}{=} \sum_{\boldsymbol{d}} \sum_{n\in\boldsymbol{d}} \sum_{i=1}^{K_T} \sum_{\substack{\tau\subseteq[K_T]:\\|\tau|=i}} (c_{n\tau} + f_{n\tau}(\boldsymbol{d})) \tag{63a}$$

$$\overset{(b)}{=} K_R \pi(N-1,K_R-1) \sum_{i=1}^{K_T} \sum_{\substack{\tau\subseteq[K_T]:\\|\tau|=i}} \sum_{n=1}^{N} (c_{n\tau} + \tilde{f}_{n\tau}) \tag{63b}$$

$$\overset{(c)}{=} K_R \pi(N-1,K_R-1) \sum_{i=1}^{K_T} b_i, \tag{63c}$$

where equality (a) holds by summing up the constraints in (50b) for all $\pi(N,K_R)$ request vectors and for all K_R files in each vector $\boldsymbol{d}$; and equalities (b) and (c) follow from the equalities in (61) and the definition of b_i in (62), respectively. From (63), we have the equality $\sum_{i=1}^{K_T} b_i = NF$.

We move on to lower bound the second term in (52a), i.e., the minimum fronthaul NDT $\delta_F^*(\boldsymbol{d})$. We start by bounding the size of the cached content. From (50c), we have

$$\begin{aligned}\mu F N K_T &\overset{(a)}{\geq} \sum_{i=1}^{K_T} \sum_{n=1}^{N} \sum_{\substack{\tau\subseteq[K_T]:\\i\in\tau}} c_{n\tau} = \sum_{n=1}^{N} \sum_{i=1}^{K_T} \sum_{\substack{\tau\subseteq[K_T]:\\i\in\tau}} c_{n\tau} \\ &\overset{(b)}{=} \sum_{n=1}^{N} \sum_{i=1}^{K_T} i \sum_{\substack{\tau\subseteq[K_T]:\\|\tau|=i}} c_{n\tau} = \sum_{i=1}^{K_T} i \sum_{\substack{\tau\subseteq[K_T]:\\|\tau|=i}} \sum_{n=1}^{N} c_{n\tau},\end{aligned} \tag{64a}$$

where inequality (a) holds by summing the inequalities in (50c) for all the K_T ENs; and equality (b) comes from the fact that the size of the cached content of a file W_n across the ENs is given as $\sum_{i=1}^{K_T} \sum_{\tau:i\in\tau} c_{n\tau} = \sum_{i=1}^{K_T} i \sum_{\tau:|\tau|=i} c_{n\tau}$.

With the above inequality, the minimum fronthaul NDT can be bounded as

$$\frac{1}{\pi(N,K_R)} \sum_{\boldsymbol{d}} \delta_F^*(\boldsymbol{d}) \overset{(a)}{\geq} \frac{1}{\pi(N,K_R)} \sum_{\boldsymbol{d}} \frac{1}{K_T} \sum_{i=1}^{K_T} \frac{1}{Fr} \sum_{n\in\boldsymbol{d}} \sum_{\substack{\tau\subseteq[K_T]:\\i\in\tau}} f_{n\tau}(\boldsymbol{d}) \tag{65a}$$

$$\overset{(b)}{=} \frac{1}{\pi(N,K_R)} \frac{1}{K_T F r} \sum_{\boldsymbol{d}} \sum_{n\in\boldsymbol{d}} \sum_{i=1}^{K_T} i \sum_{\substack{\tau\subseteq[K_T]:\\|\tau|=i}} f_{n\tau}(\boldsymbol{d}) \tag{65b}$$

$$\overset{(c)}{=} \frac{K_R}{N K_T F r} \sum_{i=1}^{K_T} i \sum_{\substack{\tau\subseteq[K_T]:\\|\tau|=i}} \sum_{n=1}^{N} \tilde{f}_{n\tau} \tag{65c}$$

$$\overset{(d)}{=} \frac{K_R}{N K_T F r} \sum_{i=1}^{K_T} i \left(b_i - \sum_{\substack{\tau\subseteq[K_T]:\\|\tau|=i}} \sum_{n=1}^{N} c_{n\tau} \right) \tag{65d}$$

$$\overset{(e)}{\geq} \frac{K_R}{K_T r} \left(\frac{1}{NF} \sum_{i=1}^{K_T} i b_i - \mu K_T \right), \tag{65e}$$

where inequality (a) holds by averaging the constraints in (50d); equality (b) follows in a manner similar to equality (b) in (64a); equalities (c) and (d) follow the equality in (61b) and the definition of b_i in (62), respectively; and inequality (e) holds by using (64a).

Now we can bound the minimum NDT by using (60f), (63) and (65a) as

$$\frac{1}{\pi(N,K_R)}\sum_{\boldsymbol{d}}\delta_E^*\big(\{c_{n\tau}\},\{f_{n\tau}(\boldsymbol{d})\},\boldsymbol{d}\big)+\delta_F^*(\boldsymbol{d}) \tag{66a}$$

$$\geq \frac{K_R}{NFn_T}\frac{(NF)^2}{\sum_{i=1}^{K_T} ib_i}+\frac{K_R}{K_Tr}\left(\frac{1}{NF}\sum_{i=1}^{K_T} ib_i-\mu K_T\right) \tag{66b}$$

$$= \frac{K_R(x-\mu K_T)}{K_Tr}+\frac{K_R}{n_T}\frac{1}{x}, \tag{66c}$$

where in the last step, we have defined the variable $x=\sum_{i=1}^{K_T} ib_i/(NF)$. Since, by (62), the expression $\sum_{i=1}^{K_T} ib_i$ is the overall number of packets of all library files that are available upon fronthaul transmission at subsets of ENs of any size i, the variable x can be interpreted as the average multiplicity of each file at the ENs after fronthaul transmission.

From (66c), we define the function

$$f(x)=\frac{K_R(x-\mu K_T)}{K_Tr}+\frac{K_R}{n_T}\frac{1}{x}. \tag{67}$$

To complete the proof, we now minimize $f(x)$ in (67) over x. To this end, we first focus on defining the domain of x. From (50f) and (65a), we have the bounds $K_R(x-\mu K_T)/(K_Tr)\leq \delta_F(\boldsymbol{d})\leq \delta_{Fmax}$, yielding the upper bound $x\leq (m_{max}-\mu K)^+ + \mu K_T = \max\{m_{max},\mu K_T\}$. We also have the inequality $x\geq \mu K_T$ due to the bound $\delta_F^*(\boldsymbol{d})\geq 0$. Furthermore, from (63), we have the inequality $\sum_{i=1}^{K_T} b_i/NF\geq 1$, yielding $x=\sum_{i=1}^{K_T} ib_i/NF\geq 1$. In summary, variable x needs to lie in the interval $\max\{1,\mu K_T\}=x_{min}\leq x\leq x_{max}=\max\{m_{max},\mu K_T\}$. We then turn to minimizing the function $f(x)$ in the interval $x\in[x_{min},x_{max}]$. Function $f(x)$ is convex for $x>0$, and the only stationary point is $x=\sqrt{K_Tr/n_T}$, i.e., $f'\big(\sqrt{K_Tr/n_T}\big)=0$. Therefore, the desired minimum f_{min} is given as

$$f_{min}=\begin{cases} f\big(\sqrt{K_Tr/n_T}\big), & \text{if } x_{min}\leq\sqrt{K_Tr/n_T}\leq x_{max}\\ \min\{f(x_{min}),f(x_{max})\}, & \text{otherwise,}\end{cases} \tag{68}$$

which is as reported in (59).

APPENDIX D
PROOF OF PROPOSITION 6

To prove Proposition 6, we first derive a lower bound $\delta'_{lb}(\mu,r)$, which is looser than the lower bound $\delta_{lb}(\mu,r)$ in Proposition 2 but more tractable. The bound leverages Proposition 4, Proposition 5 and the convexity of the minimum NDT $\delta^*(\mu,r)$ as stated in Lemma 1. The lower bounds $\delta_{lb}(\mu,r)$ and $\delta'_{lb}(\mu,r)$ are illustrated in Fig. 15.

Lemma 3: For any $r\in[0,1]$, and μ with $\mu K_T\leq m_{max}$, we have $\delta'_{lb}(\mu,r)\leq\delta_{lb}(\mu,r)$, where $\delta_{lb}(\mu,r)$ is given in (31) and we have defined

$$\delta'_{lb}(\mu,r)=\frac{(i+2-\mu K_T)K_R}{(i+1)n_T}+\frac{(\mu K_T-i-1)K_R}{(i+2)n_T} \tag{69}$$

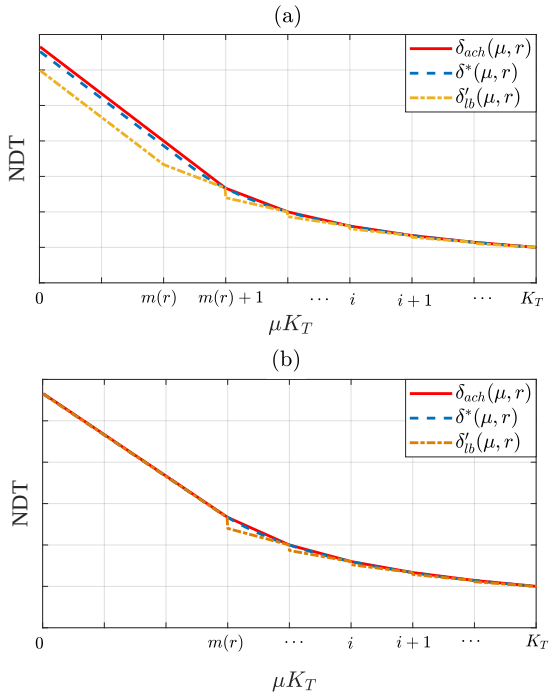

Fig. 15. Achievable NDT $\delta_{ach}(\mu, r)$ and lower bounds $\delta_{lb}(\mu, r)$ and $\delta'_{lb}(\mu, r)$: plot (a) shows Case 1 (70a), plot (b) shows Case 2 (70b).

for $\mu K_T \in [i, i+1)$, with $m(r) \le i \le m_{max} - 1$; and

$$\delta'_{lb}(\mu, r) = \begin{cases} \frac{K_R(m^*(r) - \mu K_T)}{K_T r} + \frac{K_R}{m^*(r) n_T}, & \text{if } \frac{K_T r}{n_T} \in [(m(r) - 0.5)^2, m^2(r)] \quad (70a) \\ \frac{K_R(m(r) - \mu K_T)}{K_T r} + \delta'_{lb}(\frac{m(r)}{K_T}, r), & \text{if } \frac{K_T r}{n_T} \in [m^2(r), (m(r) + 0.5)^2] \quad (70b) \end{cases}$$

for $\mu K_T \le m(r)$, where $m^*(r)$ is given in (30).

Proof: From Proposition 5, we have the equality $\delta^*(\mu, r) = \delta_{ach}(\mu, r)$ for $\mu K_T \in \{m(r)+1, \cdots, m_{max}\}$. Furthermore, we know that the minimum NDT $\delta^*(\mu, r)$ is a convex function of μ for any $r \ge 0$. Define $g_r(\mu)$ as a subgradient of the minimum NDT $\delta^*(\mu, r)$ at $\mu \in [0, 1]$ for a fixed value of r. Consider any two points μ_1 and μ_2, where $\mu_1 K_T \in \{m(r)+1, \cdots, m_{max}\}$ and μ_2 is arbitrary. By a known convex property of convex functions (see [29]), we have the inequality

$$\delta^*(\mu_2, r) \ge g_r(\mu_1)(\mu_2 - \mu_1) K_T + \delta^*(\mu_1, r). \quad (71)$$

Therefore, choosing μ_2 so that $\mu_2 K_T = \mu_1 K_T + 1$ in (71) yields

$$g_r(\mu_1) \le \delta^*(\mu_1 + 1/K_T, r) - \delta^*(\mu_1, r). \quad (72)$$

For any sub-interval $\mu K_T \in [i, i+1)$, with $m(r) \le i \le m_{max} - 1$, by setting $\mu_1 = (i+1)/K_T$ in (72), we have the bound $g_r((i+1)/K_T) \le g_{max} \triangleq \delta^*((i+2)/K_T, r) - \delta^*((i+1)/K_T, r)$. Combining with (71) and setting $\mu_2 = \mu$, we have the inequality $\delta^*(\mu, r) \ge g_{max} \cdot (\mu K_T - i - 1) + \delta^*((i+1)/K_T, r)$, which gives (69) by Lemma 1.

For the remaining interval $\mu K_T \le m(r)$, we distinguish the two cases illustrated in Fig. 15(a) and (b).

Case 1: $K_T r / n_T \in [(m(r) - 0.5)^2, m(r)^2]$. By (21) and (30) in this range, we have the inequality $m^*(r) = \sqrt{K_T r / n_T} \le$

$m(r)$. It can be directly verified that $\delta'_{lb}(\mu, r)$ in (70a) is no larger than $\delta_{lb}(\mu, r)$ in (31a) for $\mu K_T \leq m^*(r)$. Instead, for $\mu K_T \in [m^*(r), m(r)]$, since both $\delta'_{lb}(\mu, r)$ in (70a) and $K_R/(\mu K_T n_T)$ are decreasing functions of μ, they are equal for $\mu K_T = m^*(r)$, and the former has a smaller gradient for the whole range of value of μ at hand, we have $\delta'_{lb}(\mu, r) \leq K_R/(\mu K_T n_T)$, which implies that $\delta'_{lb}(\mu, r) \leq \delta^*(\mu, r)$ in (31b), as illustrated in Fig. 15(a).

Case 2: $K_T r/n_T \in [m(r)^2, (m(r)+0.5)^2]$. By (21) and (30) in this range, we have the inequality $m^*(r) = \sqrt{K_T r/n_T} \geq m(r)$. By setting $\mu_1 = (m(r)+1)/K_T$ in (72), we have $g_r((m(r)+1)/K_T) \leq g'_{max} \triangleq \delta^*((m(r)+2)/K_T, r) - \delta^*((m(r)+1)/K_T, r)$. Combining with (71) and setting $\mu_2 = m(r)/K_T$, we have the inequality $\delta^*(m(r)/K_T, r) \geq -g'_{max} + \delta^*((m(r)+1)/K_T, r)$, which gives the lower bound $\delta'_{lb}(m(r)/K_T, r)$. It is easy to verify the inequality $\delta'_{lb}(m(r)/K_T, r) \leq \delta_{lb}(m(r)/K_T, r)$. Combining this with the fact that $\delta'_{lb}(\mu, r)$ in (70b) and $\delta_{lb}(\mu, r)$ in (31a) are linear and parallel for $\mu K_T \leq m(r)$, we have $\delta'_{lb}(m(r)/K_T, r) \leq \delta_{lb}(\mu, K_T)$ in this range (see Fig. 15(b)). This completes the proof. ■

Using the lower bound $\delta'_{lb}(\mu, r)$, we can now directly compute the gap between the achievable NDT $\delta_{ach}(\mu, r)$ in Proposition 1 and the minimum NDT $\delta^*(\mu, r)$. Specifically, for $\mu K_T \in [i, i+1)$, with $m(r) \leq i \leq m_{max} - 1$, from (27) and (69), we verify that

$$\frac{\delta_{ach}(\mu, r)}{\delta'_{lb}(\mu, r)} \overset{(a)}{\leq} \frac{\delta_{ach}(\mu = i/K_T, r)}{\delta'_{lb}(\mu = i/K_T, r)} = 1 + \frac{2}{i+3i} \leq \frac{3}{2}, \tag{73}$$

where inequality (a) holds because $\delta_{ach}(\mu, r)$ and $\delta'_{lb}(\mu, r)$ are both linearly decreasing and they coincide at the endpoint $\mu K_T = i+1$. For $\mu K_T \leq m(r)$ in Case 1, from (27) and (70a), the gap is given as

$$\frac{\delta_{ach}(\mu, r)}{\delta'_{lb}(\mu, r)} \overset{(a)}{\leq} \frac{\delta_{ach}(\mu = m(r)/K_T, r)}{\delta'_{lb}(\mu = m(r)/K_T, r)} \tag{74a}$$

$$= \frac{1/m(r)n_T}{(m^*(r) - m(r))/(K_T r) + 1/(m^*(r)n_T)} \tag{74b}$$

$$\overset{(b)}{\leq} \frac{m(r)}{\sqrt{m(r)(m(r)-1)}} \overset{(c)}{\leq} \sqrt{2}, \tag{74c}$$

where inequality (a) holds because $\delta_{ach}(\mu, r)$ and $\delta'_{lb}(\mu, r)$ decrease with the same slope and the maximum ratio is at the endpoint $\mu K_T = m(r)$; inequality (b) holds due to the constraints $m^*(r) \in [\sqrt{m(r)(m(r)-1)}, m(r)]$; and inequality (c) holds for any $m(r) \geq 2$, while for $m(r) = 1$, we have $K_T r/n_T \in [0, 1]$ and $\mu \in [0, 1]$, it has been proved that $\delta_{ach}(\mu, r)$ is optimal in Proposition 5. Finally, for $\mu K_T \leq m(r)$ in Case 2, from (27) and (70b), the gap is given as

$$\frac{\delta_{ach}(\mu, r)}{\delta'_{lb}(\mu, r)} \overset{(a)}{\leq} \frac{\delta_{ach}(\mu = m(r)/K_T, r)}{\delta'_{lb}(\mu = m(r)/K_T, r)} = 1 + \frac{2}{m^2(r) + 3m(r)} \leq \frac{3}{2}, \tag{75}$$

where inequality (a) holds as inequality (a) in (74a), completing the proof.

APPENDIX E
PROOF OF PROPOSITION 8

Any achievable pipelined policy $\{\mathcal{C}_i, \{\mathcal{F}_i(b)\}_{b\in[B]}, \{\boldsymbol{v}_{inf}(b)\}_{n\in[N],f\in[F],b\in[B]}\}_{i=1}^{K_T}$ can be converted into a serial policy with parameters $\{\mathcal{C}_i, \mathcal{F}_i, \{\boldsymbol{v}_{inf}(b)\}_{n\in[N],f\in[F],b\in[B]}\}_{i=1}^{K_T}$, where $\mathcal{F}_i = \{\mathcal{F}_i(b)\}_{b\in[B]}$. In words, in the serial policy, all fronthaul

transmission takes place prior to edge communications. Using the definitions in (34)-(35), the fronthaul and edge latencies of the serial policy are given by $T_F = \sum_{b=1}^{B} T_F(b)$ and $T_E = \sum_{b=1}^{B} T_E(b)$. Furthermore, by the definition (36), we have the inequalities

$$T_P = \sum_{b=1}^{B} \max\{T_E(b), T_F(b)\} \geq \max\Big\{ \sum_{b=1}^{B} T_E(b), \sum_{b=1}^{B} T_F(b) \Big\} \geq \max\{T_E, T_F\}. \tag{76}$$

Finally, from (76), we have the inequality

$$\delta_P \geq \max\{\delta_E, \delta_F\}, \tag{77}$$

where δ_F and δ_E are the fronthaul and edge NDTs of the discussed serial policy. As a summary, any achievable pipelined NDT is lowered bounded by the maximum of the fronthaul and edge NDTs of the converted serial policy.

Recall that, in Appendix C, the fronthaul and edge NDTs under any serial policy are found to be lower bounded as (65) and (60), respectively, which yields the inequalities

$$\delta_E \geq \frac{K_R}{n_T x} \text{ and } \delta_F \geq \frac{K_R(x - \mu K_T)}{K_T r}, \tag{78}$$

where $x = \sum_{i=1}^{K_T} i b_i/(NF)$ takes values in the interval $x \in [x_{min}, x_{max}]$ with $x_{min} = \max\{1, \mu K_T\}$ and $x_{max} = \max\{m_{max}, \mu K_T\}$. To proceed, we define the two functions

$$f_1(x) = \frac{K_R}{n_T x} \text{ and } f_2(x) = \frac{K_R(x - \mu K_T)}{K_T r}. \tag{79}$$

As a result, from (77), (78) and (79), the minimum pipelined NDT δ_P^* can be bounded as

$$\delta_P^* \geq \max\{f_1(x), f_2(x)\}. \tag{80}$$

To complete the proof, we now minimize the function $f(x) = \max\{f_1(x), f_2(x)\}$ in the interval $x \in [x_{min}, x_{max}]$. Function $f(x)$ is convex for $x > 0$, and the only point whose subdifferential $\partial f(x)$ includes 0 is $x = m_p^*$, i.e., $0 \in \partial f(m_p^*)$. Hence, from [29], the minimum value $f_{p,min}$ of $f(x)$ is given as

$$f_{p,min} = \begin{cases} f(m_p^*), & \text{if } x_{min} \leq m_p^* \leq x_{max} \\ \min\{f(x_{min}), f(x_{max})\}, & \text{otherwise,} \end{cases} \tag{81}$$

which equals to $\delta_{p,lb}(\mu, r)$ in (43).

APPENDIX F
PROOF OF PROPOSITION 10

We now bound the gap between the achievable pipelined NDT $\delta_{p,ach}(\mu, r)$ in (41) and the minimum NDT $\delta_p^*(\mu, r)$ by using the lower bound $\delta_{p,lb}(\mu, r)$ in (43). There are two cases in terms of μK_T. When $\mu K_T \leq 1 - K_T r/n_T$, we have the equality $\delta_{p,ach}(\mu, r) = \delta_{p,lb}(\mu, r)$ by comparison, indicating that the achievable NDT is optimal and the gap is 1.

We move to the case $\mu K_T \geq 1 - K_T r/n_T$, which corresponds to $m_p^* \geq 1$ in (39). Directly from (40) and (42), we can obtain the multiplicities $m_p(\mu, r) = \min\{\lfloor m_p^* \rfloor, m_{max}\}$ and $m_p^*(\mu, r) = \min\{m_p^*, m_{max}\}$, respectively. By comparison, we have the inequality $m_p(\mu, r) \leq m_p^*(\mu, r)$. As a result, the gap can be bounded as

$$\frac{\delta_{p,ach}(\mu, r)}{\delta_{p,lb}(\mu, r)} \overset{(a)}{\leq} \frac{\max\left\{\frac{K_R}{m_p(\mu,r)n_T}, 1\right\}}{\max\left\{\frac{K_R}{m_p^*(\mu,r)n_T}, 1\right\}} \overset{(b)}{\leq} \frac{m_p^*(\mu, r)}{m_p(\mu, r)} \overset{(c)}{\leq} \frac{m_p^*}{\lfloor m_p^* \rfloor} \overset{(d)}{\leq} 2, \tag{82}$$

where inequality (a) holds because $\delta_{p,ach}(\mu, r)$ is a lower convex envelope of $\max\left\{K_R/(m_p(\mu, r)n_T), 1\right\}$; inequality (b) holds by considering the three different cases: $K_R/m_p(\mu, r)n_T \leq 1$, $K_R/m_p^*(\mu, r)n_T \geq 1$, and $K_R/m_p^*(\mu, r)n_T \leq 1 \leq K_R/m_p(\mu, r)n_T$, respectively, along with the fact that $K_R/(m_p(\mu, r)n_T) \geq K_R/(m_p^*(\mu, r)n_T)$; inequality (c) holds by considering the three cases: $m_p^* \leq m_{max}$, $\lfloor m_p^* \rfloor \geq m_{max}$, and $\lfloor m_p^* \rfloor \leq m_{max} \leq m_p^*$; and inequality (d) holds because $m_p^* \geq 1$. This completes the proof.

ACKNOWLEDGEMENTS

Jingjing Zhang and Osvaldo Simeone have received funding from the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme (Grant Agreement No. 725731). The authors would like to thank Roy Karasik for useful comments.

REFERENCES

[1] K. Shanmugam, N. Golrezaei, A. G. Dimakis, A. F. Molisch, and G. Caire, "Femtocaching: Wireless content delivery through distributed caching helpers," vol. 59, no. 12, pp. 8402–8413, Dec 2013.

[2] S. Traverso, M. Ahmed, M. Garetto, P. Giaccone, E. Leonardi, and S. Niccolini, "Temporal locality in today's content caching: Why it matters and how to model it," *ACM SIGCOMM*, vol. 43, no. 5, pp. 5–12, Nov. 2013.

[3] M. A. Maddah-Ali and U. Niesen, "Cache-aided interference channels," in *Proc. IEEE Int. Symp. Information Theory (ISIT)*, June 2015, pp. 809–813.

[4] A. Liu and V. Lau, "Cache-induced opportunistic MIMO cooperation: A new paradigm for future wireless content access networks," in *Proc. IEEE Int. Symp. Information Theory (ISIT)*, June 2014, pp. 46–50.

[5] N. Naderializadeh, M. A. Maddah-Ali, and A. S. Avestimehr, "Fundamental limits of cache-aided interference management," *IEEE Trans. Inf. Theory*, vol. 63, no. 5, pp. 3092–3107, May 2017.

[6] R. Tandon and O. Simeone, "Cloud-aided wireless networks with edge caching: Fundamental latency trade-offs in fog radio access networks," in *Proc. IEEE Int. Symp. Information Theory (ISIT)*, July 2016, pp. 2029–2033.

[7] A. Sengupta, R. Tandon, and O. Simeone, "Fog-aided wireless networks for content delivery: Fundamental latency tradeoffs," *IEEE Trans. Inf. Theory*, vol. 63, no. 10, pp. 6650–6678, Oct 2017.

[8] T. Q. Quek, M. Peng, O. Simeone, and W. Yu, *Cloud Radio Access Networks: Principles, Technologies, and Applications*. Cambridge University Press, 2017.

[9] Y. Cao, M. Tao, F. Xu, and K. Liu, "Fundamental storage-latency tradeoff in cache-aided MIMO interference networks," *IEEE Transactions on Wireless Communications*, vol. 16, no. 8, pp. 5061–5076, Aug 2017.

[10] J. Hachem, U. Niesen, and S. N. Diggavi, "Degrees of freedom of cache-aided wireless interference networks," *CoRR*, vol. abs/1606.03175, 2016. [Online]. Available: http://arxiv.org/abs/1606.03175

[11] J. S. P. Roig, D. Gündüz, and F. Tosato, "Interference networks with caches at both ends," in *Proc. IEEE Int. Conf. Communications (ICC)*, May 2017, pp. 1–6.

[12] N. Naderializadeh, M. A. Maddah-Ali, and A. S. Avestimehr, "Cache-aided interference management in wireless cellular networks," in *Proc. IEEE Int. Conf. Communications (ICC)*, May 2017, pp. 1–7.

[13] A. Sengupta, R. Tandon, and O. Simeone, "Cache aided wireless networks: Tradeoffs between storage and latency," in *Proc. IEEE Information Science and Systems (CISS) (ICC)*, March 2016, pp. 320–325.

[14] J. S. P. Roig, S. A. Motahari, F. Tosato, and D. Gündüz, "Fundamental limits of latency in a cache-aided 4×4 interference channel," in *Proc. IEEE Information Theory Workshop (ITW)*, Nov 2017, pp. 16–20.

[15] A. M. Girgis, O. Ercetin, M. Nafie, and T. ElBatt, "Decentralized coded caching in wireless networks: Trade-off between storage and latency," in *Proc. IEEE Int. Symp. Information Theory (ISIT)*, June 2017, pp. 2443–2447.

[16] X. Yi and G. Caire, "Topological coded caching," in *Proc. IEEE Int. Symp. Information Theory (ISIT)*, July 2016, pp. 2039–2043.

[17] N. Mital, D. Gündüz, and C. Ling, "Coded caching in a multi-server system with random topology," *CoRR*, vol. abs/1712.00649, 2017. [Online]. Available: https://arxiv.org/abs/1712.00649

[18] J. Kakar, A. Alameer, A. Chaaban, A. Sezgin, and A. Paulraj, "Cache-assisted broadcast-relay wireless networks: A delivery-time cache-memory tradeoff," *CoRR*, vol. abs/1803.04058, 2018. [Online]. Available: https://arxiv.org/abs/1803.04058

[19] J. Zhang and P. Elia, "Fundamental limits of cache-aided wireless BC: interplay of coded-caching and CSIT feedback," *IEEE Trans. Inf. Theory*, vol. 63, no. 5, pp. 3142–3160, May 2017.

[20] J. Kakar, S. Gherekhloo, Z. H. Awan, and A. Sezgin, "Fundamental limits on latency in cloud- and cache-aided hetnets," in *Proc. IEEE Int. Conf. Communications (ICC)*, May 2017, pp. 1–6.

[21] J. Goseling, O. Simeone, and P. Popovski, "Delivery latency trade-offs of heterogeneous contents in fog radio access networks," in *Proc. IEEE Global Conf. Communications (GLOBECOM)*, December 2017.

[22] S. M. Azimi, O. Simeone, A. Sengupta, and R. Tandon, "Online edge caching and wireless delivery in fog-aided networks with dynamic content popularity," *CoRR*, vol. abs/1711.10430, 2017. [Online]. Available: https://arxiv.org/abs/1711.10430

[23] S. M. Azimi, O. Simeone, and R. Tandon, "Content delivery in fog-aided small-cell systems with offline and online caching: An information-theoretic analysis," *Entropy*, vol. 19, no. 7, 366, 2017.

[24] J. S. P. Roig, F. Tosato, and D. Gündüz, "Storage-latency trade-off in cache-aided fog radio access networks," *CoRR*, vol. abs/1802.01983, 2018. [Online]. Available: https://arxiv.org/abs/1802.01983

[25] S. M. Azimi, O. Simeone, A. Sengupta, and R. Tandon, "Online edge caching in fog-aided wireless network," *CoRR*, vol. abs/1701.06188, 2017. [Online]. Available: http://arxiv.org/abs/1701.06188

[26] D. R. Stinson, *Combinatorial Designs: Construction and Analysis*. Springer, 2004.

[27] W.-T. Chang, R. Tandon, and O. Simeone, "Cache-aided content delivery in Fog-RAN systems with topological information and no CSI," *Proc. Asilomar Conf. Signals, Systems and Computers*, Nov. 2017.

[28] V. Bioglio, F. Gabry, and I. Land, "Optimizing MDS codes for caching at the edge," *CoRR*, vol. abs/1508.05753, 2015. [Online]. Available: http://arxiv.org/abs/1508.05753

[29] S. Boyd and J. Duchi, "Lecture slides: Subgradients." [Online]. Available: https://see.stanford.edu/materials/lsocoee364b/01-subgradients_slides.pdf